Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 4, pp. 1289 ~ 1297

An Algorithm for Baseline Correction of SELDI/MALDI Mass Spectrometry Data[1]

Kyeong Eun Lee[2]

Abstract

Before other statistical data analysis the preprocessing steps should be performed adequately to have meaningful results. These steps include processes such as baseline correction, normalization, denoising, and multiple alignment. In this paper an algorithm for baseline correction is proposed with using the piecewise cubic Hermite interpolation with block-selected points and local minima after denoising for SELDI or MALDI mass spectrometry data.

Keywords : Baseline correction, Denoising, Piecewise cubic Hermite interpolation, SELDI/MALDI MS data

1. Introduction

Proteome is a term used to describe the whole complements of 'PROTEins' encoded by the 'genOME' in a given cell, tissue, or organism at a particular time, primarily coined by Mark Wilkins in 1995. Proteomics is the comprehensive study of the proteome, especially, its structures, post-translational modifications, interactions and functions. Since, dissimilar to genome, proteome is different from cell to cell and changes through its interactions with the genome or the environment, proteome research can be more helpful to examine the direct causes of diseases than other genome researches, by the qualitative or quantitative comparison of proteomes under different conditions.

On the other hand, there are various characterization methods for proteins, because proteins have much more complex structure than DNA. However, high-throughput techniques of protein identification or quantification among the

[1] This research was supported by Kyungpook National University Research Fund, 2005.

[2] Full-time Instructor, Dept. of Statistics, Kyungpook National University, Daegu. 702-701, Korea. E-mail: artlee@knu.ac.kr

branches of proteomics are mainly based on mass-spectrometry techniques, such as, Matrix Assisted Laser Desorption/Ionization (MALDI) ion source, and Time of Flight (TOF) detection system, and Surface Enhanced Laser Desorption/Ionization (SELDI)-TOF.

The main purpose of cancer study with mass-spectrometry analyses is to identify distinctive proteins between in samples of cancer patients and in those of normal patients, and further to examine the biological processes and path ways. However, before these full scale data analyses, it is necessary to preprocess the mass spectrometry data, such as baseline correction, normalization, denoising, and multiple alignment. In this paper, we focus on baseline correction. There are continuous efforts to develop better methods for baseline corrections, including several recent efforts: local polynomial regression using weighted least squares, local linear regression (Wu et al, 2003), a semi-monotonic baseline correction (Baggerly et al., 2003), nonlinear filter known as the tophat operator (Sauve and Speed, 2004), and Heuristic-based baseline removal algorithm (Lin et al., 2006).

Some baseline correction methods are implemented on raw noisy mass spectrometry data, but our algorithm is based on the denoised mass spectrometry since we believe that the baseline should not be affected by random noise. After the stationary wavelet transform (SWT), Ebayes thresholds (Johnstone and Silverman, 2005) are applied to each level of the transform.

Since the SELDI/MALDI baselines are known as a smooth and downward drifting curve moving from low m/z and to high m/z, and sparse signals are added to the baselines, the estimated baseline should be very smooth and lower than the observed spectrum and the baseline corrected signal should be flat on non-peak regions. In this paper, we develop such an algorithm for baseline correction using the piecewise cubic Hermite interpolation, which preserves monotonicity and the shape of the data, based on a set of time points with some chosen local minima and flat blocks.

2. Background

2.1 SELDI Mass Spectrometry Data

Since SELDI-MS overcomes some of the problems associated with sample preparation inherent in MALDI-MS (Bensamail, et al., 2005) and MALDI baselines share characteristics with SELDI baselines, we briefly mention only SELDI-MS in this section. SELDI-MS data are generated as follows: Samples, such as serum or tissue, are directly applied to the surfaces (stainless steel or aluminium based supports, or chips, engineered with chemical or biological bait surfaces). Samples are then washed to remove non-specifically or weekly bound proteins or buffer contaminants. Energy absorbing molecules (EAM) are added to retained proteins to promote laser-based desorption and ionization. The proteins and EAM are co-crystallized and bombarded with a pulsed-UV laser beam, causing them to vaporize and ionize. The gaseous ionized proteins are accelerated in an electric field and strike a detector. Their mass-to-charge is obtained by their time-of-flight (Wulfkuhle et al., 2003).

The baseline is a mass-to-charge dependent offset, due to the chemical EAM matrix, which makes difficult to compare with different spectra since it varies between different samples. Therefore, after baseline correction, the comparisons between intensities of m/z values of different spectra can be meaningful.

2.2 The Stationary Wavelet Transform and Ebayes Threshold

Wavelet methods are known to be a powerful tool in nonparametric regression due to their adaptability to locally irregular curves (Antoniadis et al., 2001). The nonparametric regression model can be represented as

$$Y_i = f(t_i) + \epsilon_i, i = 1, \ldots, n$$

where f is the underlying unknown function at equally spaced points $\{t_1, t_2, \ldots, t_n\}$ and ϵ_i's are independent $N(0, \sigma^2)$ random errors. The discrete wavelet transform (DWT) of $\boldsymbol{f} = (f(t_1), \cdots, f(t_n))'$ and $\boldsymbol{y} = (y_1, \cdots, y_n)'$ are given by $\boldsymbol{d} = W'\boldsymbol{f}$ and $\boldsymbol{d}^* = W'\boldsymbol{y}$ respectively, where W is an orthogonal matrix and then the nonparametric regression model can be expressed in the wavelet domain as

$$d^*_{jk} = d_{jk} + \epsilon_{jk}, k = 1, \ldots, 2^{j-1}, j = 1, \ldots, J$$

where the double indices show the nature of multiresolution wavelet decomposition. (Clyde and George, 2000). Basically DWT has two steps: low and high filtering step and decimation step and due to the latter step, DWT is computationally fast and compact in terms of storage space. But DWT is not shift invariant: the wavelet coefficients of the shifted signal are different from those of the original signal. In order to get rid of this shift-variant property, several wavelet methods such as stationary wavelet transform (SWT), redundant wavelet transform, shift invariant transform, or undecimated wavelet transform, have been invented independently but they are intrinsically equivalent (Guo et al. 1995). "The basic idea of the SWT is to 'fill in the gaps' caused by the decimation step in the standard wavelet transform."(Nason and Silverman, 1995). And the noise reduction performance based on the SWT becomes much better than those based on the DWT (Guo et al. 1995).

Bayesian approaches model naturally sparsity of wavelet coefficients using an appropriate prior distribution. Clyde et al. (1998) and Abramovich et al. (1998) have considered a scale mixture for each wavelet coefficient

$$d_{jk} \sim (1-\pi_j)\delta(0)+\pi_j N(0,\tau_j^2), k=1,\ldots,2^{j-1}, j=1,\ldots,J. \qquad (2.1)$$

And Johnstone and Silverman (2005) used a heavy-tailed distribution instead Normal distribution in (2.1) since they agreed with Wainwright et al. (2001)'s arguments for a heavy-tailed marginal distribution of wavelet coefficients and they took the marginal likelihood estimator for the mixing weight. They took the Bayes rule corresponding to L_1-loss, the posterior median, which actually leads to a thresholding rule, so it is called as Ebayes Threshold.

2.3 Piecewise Cubic Hermite Interpolation

Given n points in the plane, (x_k, y_k), $k=1,\ldots,n$, with distinct x_k's, let $h_k = x_{k+1} - x_k$ and the first divided differences, δ_k, is given by $\delta_k = \frac{y_{k+1}-y_k}{h_k}$. Let d_k denote the slope of the interpolant at x_k, $d_k = P'(x_k)$. Consider the following function on the interval $x_k \le x \le x_{k+1}$, expressed in terms of local variables $s = x - x_k$ and $h = h_k$,

$$P(x) = \frac{3hs^2-2s^3}{h^3} y_{k+1} + \frac{h^3-3hs^2+2s^3}{h^3} y_k + \frac{s^2(s-h)}{h^2} d_{k+1} + \frac{s(s-h)^2}{h^2} d_k.$$

This is a cubic Hermite interpolation if it satisfies the following four conditions:

$$P(x_k) = y_k, P(x_{k+1}) = y_{k+1}, P'(x_k) = d_k, P'(x_{k+1}) = d_{k+1.}$$

Cubic Hermite interpolation is a shape-preserving with true continuity between the interpolants. Further details are referred to Moler (2004).

3. Baseline Correction

Coombes et al. (2005) suggested a mathematical model for mass-spectrum intensity for spectrum i at TOF t_j

$$y_i(t_j) = B_i(t_j) + N_i S_i(t_j) + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2(t_j))$$

where $B_i(t)$ is a baseline, $S_i(t)$ is a true signal, N_i is a normalizing factor, $\ni_{ij}$

is a random error. Since each spectrum has its own baseline, if the baseline is not satisfactorily corrected, a direct comparison between spectra is not meaningful. In this paper, we suggest an algorithm for denoising and baseline correction. If it works well, residuals should be randomly dispersed and the denoised and baseline corrected signal should have flat and zero height non-peak areas.

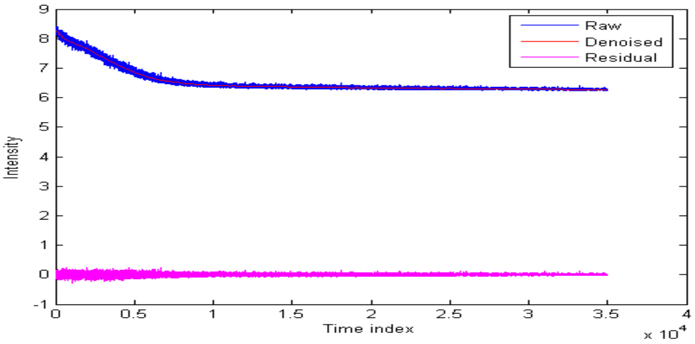
The key ideas of our algorithm are as follows:

1. In denoising the noise spectrum, SWT was utilized by taking the SWT of the spectrum, applying the empirical Bayes method separately to each level of the transform, applying the average basis inverse of the resulting SWT denoised spectrum, then, getting the residual (or estimated error) function by subtraction denoised spectrum from the original spectrum.
2. Find local minima by a moving window. And deselect some local minima in a cluster of peaks by checking their slopes.
3. To identify flat blocks, compare the 95th relative amplitude of signal with the block-varying threshold, such as 1.645 MAD of residuals in each block whose endpoints are in the selected set of time points in the previous step. All time points in those blocks with relatively small amplitudes are identified flat blocks.
4. To estimate a baseline, the piecewise cubic interpolation was used with points in step 3 (local minima) and step 4 (flat blocks).
5. The baseline corrected signal is obtained by subtraction the estimated baseline from the denoised signal.

4. Examples

We applied this method for finding a good baseline to two data sets: Pawitan Blank SELDI-MS dataset (Tan et al., 2006) and a published MALDI-MS data set (Wang et al., 2003).

Pawitan blank dataset is obtained from Ciphergen SAX2 chips using buffer only (Figure 1). So it has a zero protein signal with baseline and noise. After denoising, if there were no local minima in a too broad interval, baselines tend to show the monotonically decreasing phenomena. Therefore, it is preferred to find the local minima in segmented smaller intervals of the broad interval. In this blank data set, the relative amplitude of signal at each block is not bigger than our block-varying threshold, therefore, all time points are considered as flat regions(Figure 2 Top).

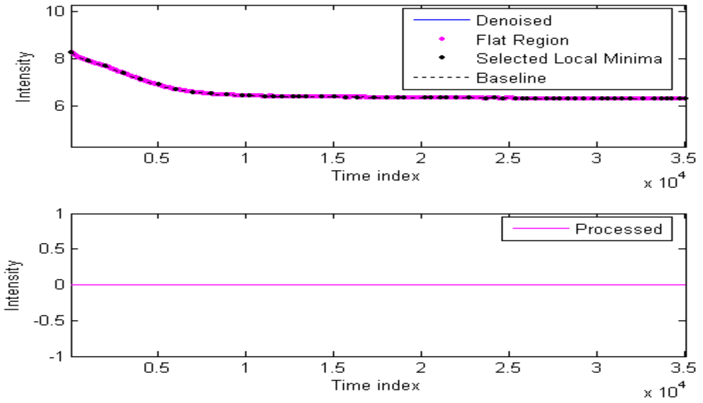


<Figure 1> Pawitan Blank Data and Residual after Denoising

We estimated the signal using our algorithm by

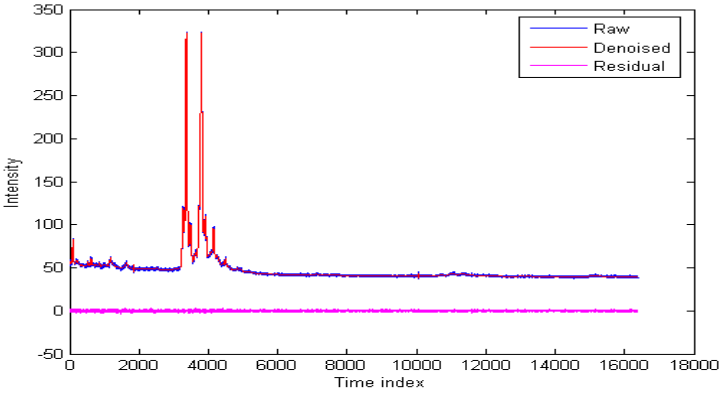
$$\widehat{S}(t)=\widehat{y}(t)-\widehat{B}(t)$$

and got the zero signal (Figure 2 Bottom).

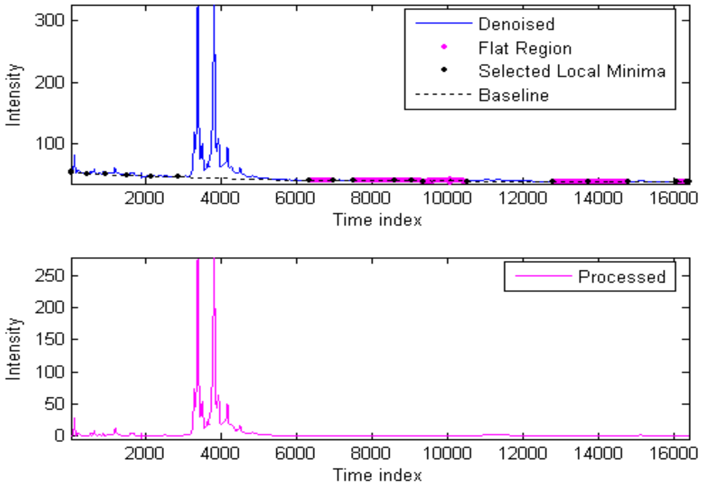


<Figure 2> Selected Time Points (local minima and flat regions) and the Estimated Signal

The second data set is a published MALDI-MS data(Wang et al., 2003). In general, SELDI or MALDI-MS data has non-flat baseline in low masses and almost flat or non-zero baseline in high masses(Figure 3).



<Figure 3> Raw MS Data and Residuals after Denoising



<Figure 4> Selected Time Points (local minima and flat regions) and the Estimated Signal.

After the SW, the denoised MS (red solid line in Figure 3) is obtained and then the local minima by a moving window are found and the local minima in a cluster of peaks are deselected by checking their slope (black dots in Figure 4 Top). Each interval is determined as flat or peak area by comparing the relative amplitude of signal with the block-dependent threshold, based on MAD of residuals in each block. Flat blocks are indicated by magenta dots in Figure 4 Top. The piecewise cubic interpolation with flat regions and selected local minima is used to estimate the baseline. And the estimated signal is produced by our proposed algorithm in Figure 4 Bottom.

5. Concluding Remark

The denoising and baseline correction part among the preprocessing steps are mainly dealt in this paper. Since the residuals from denoising with E-Bayes threshold after SWT can be regarded as noise, the flat area can be discerned by the comparison between the relative amplitude of the denoised MS and the threshold determined by MAD of the noise from each interval. The method of baseline correction is proposed with using the piecewise cubic interpolation, which preserves the shape with the selected points, consisting of flat area points and some selected local minima. The characteristic of our proposed method is that the signal size of the flat-assumed area is zero. Actually raw MS data sets are very noisy but have sparse peaks, so, it is reflected that the size of the signal is zero other than those peak area.

References

1. Abramovich, F., Sapatinas, T. and Silverman, B.W. (1998). Wavelet Thresholding via a Bayesian Approach. *Journal of the Royal Statistical Society. Series B*, Vol. 60, 725-749.
2. Antoniadis, A., Bigot, J. and Sapatinas, T. (2001). Wavelet Estimators in Nonparametric Regression: A Comparative Simulation Study. Journal of Statistical Software, Vol. 6, Issue 6.
3. Baggerly, K.A., Morris, J.S., Wang, J. et al. (2003). A Comprehensive Approach to the Analysis of MALDI-TOF Proteomics Spectra from Serum Samples. Proteomics, 3, 1667-1682.
4. Bensmail, H., Goleck, J., Moody, M.M. et al. (2005). A Novel Approach for Clustering Proteomics Data Using Bayesian Fast Fourier Transform. *Bioinformatics.* Vol. 21, 2210-2224.
5. Clyde, M. and Parmigiani, G. and Vidkovic, B. (1998). Multiple Shrinkage and Subset Selection in Wavelets. *Biometrika.* Vol. 85, 391-401.
6. Clyde, M. and George, E.I. (2000). Flexible empirical Bayes Estimation for wavelets. *Journal of the Royal Statistical Society. Series B*, Vol. 62, 681-698.
7. Coombes, K.R., Koomen, J.M., Baggerly, K.A. et al. (2005). Understanding the Characteristics of Mass Spectrometry Data through the Use of Simulation. *Cancer Informatics*. Vol. 1, 41-52.
8. Guo, H., Lang, M., Odegard, J.E., and Burrus, C.S. (1995). Nonlinear Shrinkage of Undecimated DWT for Noise Reduction and Data Compression. In *Proceedings of International Conference on Digital*

Signal Processing, Limassol, Cyprus, June 1995.

9.Johnstone, I.M. and Silverman, B.W. (2005). Empirical Bayes Selection of Wavelet Thresholds. *The Annals of Statistics*. Vol. 33, 1700-1752.

10. Lin, S.M., Du, P. and Kibbe, W.A. (2006). Heuristic-Based Baseline Removal Algorithm for SELDI Proteomics Data. Manuscript. Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago.

11. Moler C. (2004). Numerical Computing with Matlab. SIAM.

12. Nason, G.P. and Silverman, B.W. (1995). The stationary wavelet transform and some statistical applications. Pages 281-300 of: Antoniadis, A., & Oppenheim, G. (eds), Wavelets and Statistics, Lecture Notes in Statistics 103. New-York: Springer-Verlag.

14. Suave, A.C. and Speed, T.P. (2004) Normalization, Baseline Correction and Alignment of High-Throughput Mass Spectrometry Data. *Proceedings of the Genomics Signal Processing and Statistics Workshop*, Baltimore, MO, USA.

15. Tan, C.S., Ploner, A., Quandt, A. et al. (2006). Finding Regions of Significance in SELDI Measurements for Identifying Protein Biomarkers. *Bioinformatics.* Vol. 22, 1515-1523.

16. Wainwright, M.J., Simoncelli, E.P. and Willsky, A.S. (2001). Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Applied and Computational Harmonic Analysis*, Vol. 11, 89-123.

17. Wang, M.Z., Howard, B., Campa M.J. et al. (2003). Analysis of Human Serum Proteins by Liquid Phase Isoelectric Focusing and Matrix-Assisted Laser Desorption/Ionization-Mass Spectrometry. *Proteomics.* Vol. 3, 1661-1666.

18. Wu, B., Abbott, T., Fishman, D., et al. (2003). Comparison of Statistical Methods for Classification of Ovarian Cancer Using Mass Spectrometry Data. *Bioinformatics.* Vol. 19, 1636-1643.

19. Wulfkuhle, J.D., Liotta, L. and Petricoin, E.F. (2003). Proteomics Applications for the Early Detection of Cancer. *Nature.* Vol. 3, 267-275.Baumgarther, W., Weib, P., and Schindler, H. (1998). A Nonparametric Test for the General Two-sample Problem, *Biometrics,* 54, 1129-1135.

[received date : Oct. 2006, accepted date : Nov. 2006]