Peaks detection and alignment for mass spectrometry data

Anestis Antoniadis [1], Jérémie Bigot [2,3] and Sophie Lambert-Lacroix [1]

Titre: Détection et alignement de pics en spectrométrie de masse

Abstract: The goal of this paper is to review existing methods for protein mass spectrometry data analysis, and to present a new methodology for automatic extraction of significant peaks (biomarkers). For the pre-processing step required for data from MALDI-TOF or SELDI-TOF spectra, we use a purely nonparametric approach that combines stationary invariant wavelet transform for noise removal and penalized spline quantile regression for baseline correction. We further present a multi-scale spectra alignment technique that is based on identification of statistically significant peaks from a set of spectra. This method allows one to find common peaks in a set of spectra that can subsequently be mapped to individual proteins. This may serve as useful biomarkers in medical applications, or as individual features for further multidimensional statistical analysis. MALDI-TOF spectra obtained from serum samples are used throughout the paper to illustrate the methodology.

Résumé : Le but de cet article est de faire une revue des méthodes existantes pour l'analyse de données protéomiques issues de spectromètres de masse, et de présenter une nouvelle méthodologie pour l'extraction automatique de pics significatifs (bio-marqueurs). Pour les étapes de pré-traitement nécessaires pour des données issues de spectres MALDI-TOF ou SELDI-TOF, nous utilisons une approche purement nonparamétrique qui combine la transformée en ondelettes invariante par translation pour le débruitage et la régression quantile pénalisée à partir de splines pour la correction de la ligne de base. Nous présentons ensuite une technique d'alignement multi-échelle qui est basée sur l'identification des pics statistiquement significatifs dans un ensemble de spectres. Cette méthode permet de trouver les pics communs à un ensemble de spectres qui peuvent être associés aux protéines des individus. Ceux-ci peuvent servir de bio-marqueurs utiles pour des applications médicales, ou bien de vecteurs de caractéristiques pour une analyse statistique multi-dimensionnelle des individus. Des spectres MALDI-TOF obtenus à partir d'échantillons de sérum sont utilisés à travers tout l'article pour illustrer la méthodologie.

Keywords: nonparametric regression, wavelets, regression quantiles, landmark detection, curve alignment, biomarkers identification
Mots-clés : regression nonparamétrique, ondelettes, régression quantile, détection de pic, alignement de courbes, identification de biomarqueurs
AMS 2000 subject classifications: 62G05, 62G08, 62-02, 62-07

[1] Laboratoire Jean Kuntzmann, University Joseph Fourier, BP 53, 38041 Grenoble, France.
E-mail: Anestis.Antoniadis@imag.fr; Sophie.Lambert@imag.fr
[2] Institut de Mathématiques de Toulouse, Université de Toulouse et CNRS (UMR 5219), 31062 Toulouse, France.
E-mail: Jeremie.Bigot@math.univ-toulouse.fr
[3] Center for Mathematical Modeling, Universidad de Chile, Santiago, Chile.

1. Introduction

Microarray data has been successfully used to identify genes responsible for many diseases. However, although proteins are coded by genes, there is no one-to-one relationship between the protein and the mRNA due to different rates of translation. Hence studying mRNA expressions (microarrays) may be an indirect way to understand a disease etiology. This ineffectiveness of genomics caused a big shift of interest from genomics to proteomics. Since, the proteins are the controllers of all cell functions, they are closely connected to many diseases and metabolic processes, and thus it is hoped that proteomic studies may provide a more direct information for understanding the biological functions towards a disease profile. An important tool used for protein identification and high throughput comparative profiling of disease and non-disease complex protein samples in proteomics is mass spectrometry (MS). With this technology it is possible to identify specific biomarkers related to a given metabolic process or disease from the lower molecular weight range of the circulating proteome from easily obtained biological fluids such as plasma or serum. Recent research has demonstrated that using such technology to generate protein expression profiles from lung cancer lystates is an alternative promising strategy in the search for new diagnostic and therapeutic molecular targets.

There are at least two kinds of mass spectrometry instruments commonly applied to clinical and biological problems today, namely, Matrix-Assisted Laser Desorption and Ionization Time-Of-Flight (MALDI-TOF) and Surface-Enhanced Laser Desorption and Ionization Time-Of-Flight (SELDI-TOF) mass spectroscopy. Whatever is the technique used, one obtains from a biological sample a calibrated output which is a mass spectrum characterized by numerous peaks, which correspond to individual proteins or protein fragments (polypeptides) present in the sample. The heights of the peaks represent the intensities or abundance of ions in the sample for a specific mass to charge ratio (m/z) value. These heights along with the m/z values represent the fingerprint of the sample. Hence detecting location and amplitude of common peaks from a set of spectra is a way to identify specific biomarkers that can be used to characterize patients and to compare groups of individuals.

Such techniques result in a huge amount of data to be analyzed and generate a need for a rapid, efficient and fully automated method for comparing multiple MS spectra. Raw spectra acquired by TOF mass-spectrometers are generally a mixture of a real signal, noise of different characteristics and a varying baseline. Statistically, a possible model for a given mass spectrometry (MS) spectrum is to represent it schematically by the equation

$$Y(m/z) = B(m/z) + NS(m/z) + \varepsilon(m/z) \tag{1}$$

where $Y(m/z)$ is the observed intensity of the spectrum at mass to charge ratio m/z, $B(m/z)$ is the baseline representing a relatively smooth artifact commonly seen in mass spectrometry data, $S(m/z)$ is the true signal of interest consisting of a sum of possible overlapping peaks, N is a normalization factor to adjust for possibly differing amounts of protein in each sample, and $\varepsilon(m/z)$ is an additive white noise with variance σ_ε^2 arising from the measurement process.

In our numerical experiments, we will consider samples of nipple aspirate fluid (NAF) from breast cancer patients and from healthy women (for a complete description, see [12]). These data are available from the web site `http://bioinformatics.mdanderson.org/pubdata.html` and have been used by [12] to look at the reproducibility of their method in detecting and

identifying relevant peaks, since the 24 spectra of the NAF data were independently derived from the same starting material. Typical examples of raw spectra are displayed in Figure 1. The three spectra shown in Figure 1 will be used as an illustrative example throughout the paper.

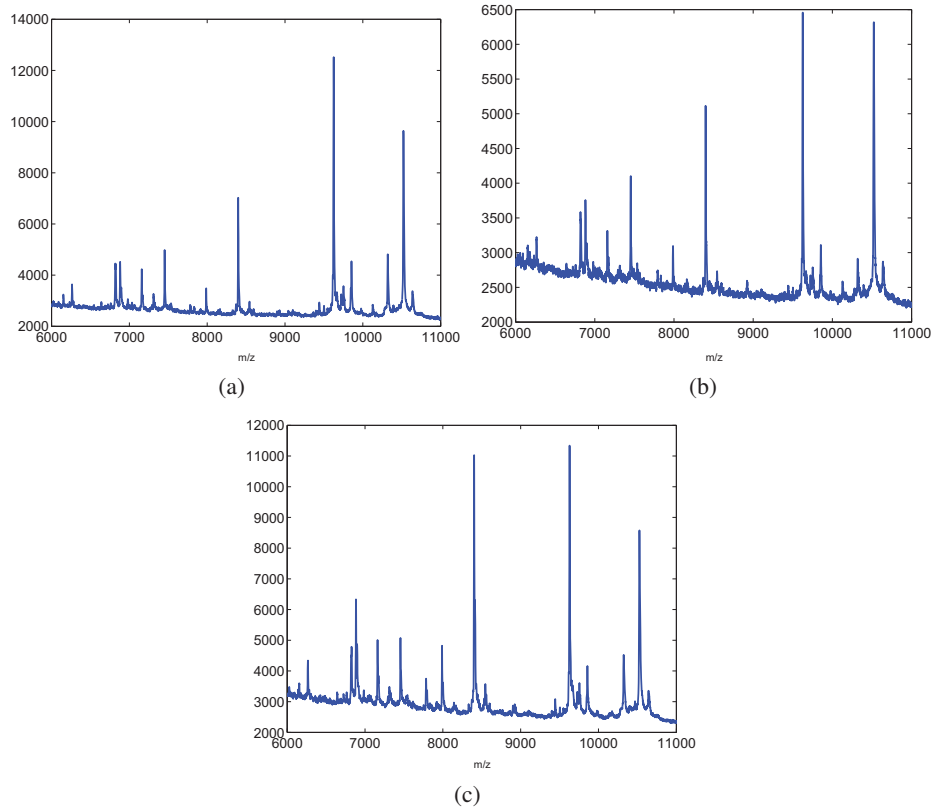

FIGURE 1. *A typical example of three raw spectra from different individuals from the NAF data used in [12]. The horizontal axis is the mass to charge ratio (m/z) in Daltons. The vertical axis is the abundance of ions in the sample for a specific m/z value.*

Pre-processing of such data is therefore extremely important to extract $S(m/z)$ which is the signal of interest. Indeed, inadequate or incorrect preprocessing methods can result in data sets that exhibit substantial biases and make it difficult to reach meaningful conclusions.

In this paper, we propose to use the several preprocessing steps that we have developed in detail in [3] before analyzing MS data. These preprocessing steps include translation invariant wavelet denoising, baseline correction and normalization along the m/z axis. These pre-processing steps allow one to obtain an estimator of the true normalized spectra $S(m/z)$. However, to compare spectra from different samples further statistical analysis has to be performed. Indeed, biologically significant comparisons for such data are all based on the alignment of spectra where the ultimate goal is to identify differentially expressed proteins in samples through the identification of common peaks. MS spectra alignment is difficult even after instrument calibration with internal markers because the mass errors vary with m/z in a nonlinear fashion as a result of experimental and instrumental complexity and data variation. In this paper, we propose to perform automatic

landmark-based curve alignment by combining the multiscale technique developed in [4, 5] to detect significant landmarks (such as locations of maxima) in a set of curves with recent results in [7, 6, 8, 9, 23] on curve and image warping. The main goal of this paper is to show that such an approach is particularly well-suited for identifying common peaks from a sample of normalized spectra. Such peaks can then be used as individual features for multidimensional statistical analysis (such as classification, clustering, ANOVA), or as specific biomarkers for biomedical applications.

The rest of the paper is organized as follows. Section 2 describes the pre-processing and normalization methods proposed in [3]. Section 3 is devoted to the presentation of a multiscale approach to automatically align and detect multiple peaks in MS spectra. Throughout the paper, we review existing methods for MS spectra analysis, and we discuss their differences or similarities with our approach.

2. Pre-processing steps

In this section, we describe the approach followed in [3] for pre-processing raw MS spectra. These results have been shown to be excellent when compared to other pre-processing methods in MS spectra analysis.

2.1. Wavelet denoising

To isolate and remove the additive noise component of a spectrum we use a translation invariant wavelet transform in a spirit similar to the Undecimated Discrete Wavelet Transforms (UDWT) filtering method of [12] for denoising SELDI-TOF spectra. One reason for the popularity of the wavelet transform (see for example [25] and [13]) is that measured data from most processes are inherently multiscale in nature. Consequently, data analysis and modeling methods that represent the measured variables at multiple scales are better suited for extracting information from measured data than methods that represent the variables at a single scale. This is why wavelets have recently received attention as a tool for preprocessing mass spectra (see e.g. [10, 27, 12]).

Briefly, the following steps are used to denoise an observed n-length mass spectrum Y: a wavelet transform is used to transform Y, certain subsets of coefficients are thresholded, then the inverse transform is applied to obtain the denoised signal.

In this paper we chose to use a translation invariant wavelet transform introduced in [11] that is based of the technique of "cycle spinning"; see also [32]. Note that the denoising performances can change drastically if one uses a wavelet transform that is not translation invariant. The idea of denoising via cycle spinning is to apply denoising not only to Y, but also to all possible unique circularly shifted versions of Y, and to average the results.

It is important to recall that for Y of length n cycle spinning is a transform that is overdetermined and produces a mean-zero wavelet coefficient sequence $\{\tilde{W}_{j,t}^{(Y)}, t = 0, \ldots, n-1\}$ at each level j of the transform. Such a sequence can be written as the length-n column vector $\tilde{\mathbf{W}}_j^{(Y)} = W_j Y$ where W_j is the level-j stationary wavelet transform matrix that maps Y to $\tilde{\mathbf{W}}_j^{(Y)}$. Then, we have adopted universal and hard thresholding for denoising MS spectra. Note that $j = 1$ corresponds to the finest resolution, and $j = J_0$ to the coarsest such that $n \approx 2^{J_0}$. Then the algorithm is:

1. Compute a level J_0 stationary wavelet transform giving coefficient vectors $\tilde{\mathbf{W}}_1^{(Y)}, \ldots, \tilde{\mathbf{W}}_{J_0}^{(Y)}$ and $\tilde{\mathbf{V}}_{J_0}^{(Y)}$, where $\tilde{\mathbf{V}}_{J_0}^{(Y)}$ denotes the vector of scaling coefficients at resolution J_0.
2. For each $j = 1, \ldots, J_0$ apply hard thresholding using the level-dependent universal threshold with $\sigma_{n_j}^2 = \sigma_\varepsilon^2 / 2^j$, to obtain

$$\hat{W}_{j,t}^{(Y)} = \begin{cases} \tilde{W}_{j,t}^{(Y)}, & \text{if } \left|\tilde{W}_{j,t}^{(Y)}\right| > \sigma_{n_j}\sqrt{2\log n} \\ 0, & \text{otherwise.} \end{cases}$$

3. The denoised signal is then obtained by applying the inverse stationary wavelet transform to $\hat{\mathbf{W}}_1^{(Y)}, \ldots, \hat{\mathbf{W}}_{J_0}^{(Y)}$ and $\tilde{\mathbf{V}}_{J_0}^{(Y)}$.

Since σ_ε is unknown, it is estimated by the MAD scale estimate defined by

$$\hat{\sigma}_{\text{MAD}} = \frac{\text{median}\left\{|\tilde{W}_{1,0}^{(Y)}|, |\tilde{W}_{1,1}^{(Y)}|, \ldots, |\tilde{W}_{1,n-1}^{(Y)}|\right\}}{0.6745}.$$

After denoising in this manner, separating background from true signal is considerably easier. Note that the universal threshold is a common choice in wavelet denoising that is nearly optimal in the case of Gaussian noise. Moreover the MAD procedure is a robust estimation technique, and the universal threshold is not restricted to a Gaussian setting.

2.2. Baseline correction

MS spectra frequently exhibit a decreasing baseline that is unrelated to constituent protein composition. Clearly, such a nuisance variation has to be removed before applying any meaningful quantitative analysis since such a background component is added to the real signal and overstates the intensities of peaks. Varying background present in warped spectra makes it difficult to properly calculate their similarity, and thus another issue caused by varying baseline is the difficulty with aligning spectra by maximizing a similarity measure between them (see [17]).

Many numerical methods have been developed for estimating the baseline component. Among these techniques are methods based on digital filters [35, 31]. Such filters usually introduce artefacts and simultaneously distort the real signal. Other approaches rely on automated peak rejection [30, 12]. These algorithms fit some functions to find regions of signal that consist only of the baseline without peaks of real signal. The functions being fitted may have different forms, e.g. polynomials, splines. The main disadvantages of peak-rejection approaches are difficulties related to identification of peak-free regions. On the other hand, threshold-based rejection of peaks gives good results when the baseline is relatively smooth [36], and fails for signals with significantly varying baseline.

Because of difficulties caused by automatic peak rejection other approaches have been designed to fit a baseline without detecting the peaks. In [1], the baseline is fitted with a low-order polynomial that prevents it from fitting the real signal peaks. For signals with many positive peaks, however, e.g. mass spectra, the baseline estimated in this way has values which are too high. Subtraction of a background with values which are too high from a signal introduces significant distortions to the analyzed signal, i.e. the values for peaks are too low. Other approaches rely on

statistical methods, such as maximum entropy [26]. There are also approaches based on baseline removal in the wavelet domain [24].

In [3] we have proposed an automated approach for background elimination based on penalized regression quantile splines. This procedure may be regarded as a method similar to the "peak rejection" approaches. However, there is no need to detect peaks. It is based on polynomial spline representation of the varying background signal and a L_1-penalization of a regression quantile loss function.

To be more precise, this method can be summarized as follows (see [3] for a more detailed presentation): denote by x the m/z variable and suppose that the baseline function denoted by $B(\cdot)$ is sufficiently regular to be well represented by a linear combination of known basis functions $\{h_j,\ j = 1,\ldots,K\}$,

$$B(x) = \sum_{j=1}^{K} \beta_j h_j(x),$$

Then, estimating the baseline amounts to estimate the coefficients $\beta = (\beta_1,\ldots,\beta_K)^T$. As an example of such basis functions we took polynomial splines (basic references are [14] and [15]). Polynomial regression splines of degree p with knots $\tilde{x}_1 < \cdots < \tilde{x}_q$ may be represented by a family of piecewise polynomial functions $\{h_j, j = 1,\ldots,K\}$ with $K = p+q+1$. Assuming the initial location of the knots known, the K dimensional parameter vector β describes the polynomial coefficients to represent the function B. The baseline can then be written as $B(x) = H(x)\beta$ where, for x given, $H(x)$ is the matrix whose columns are $h_j(x)$, for $j = 1,\ldots,K$. Usually the number K of basis functions used in the representation of B should be large in order to give a fairly flexible way for approximating B. A small K may result in a function space which is not flexible enough to capture the variability of the baseline, but a too large number of basis functions may lead to serious overfitting and as a consequence to an underestimation of the intensities of the peaks. To automatically select a small number of basis functions, we consider the following strategy: an initial fixed large number of potential knots is chosen at fixed quantiles of the x variable with the intention to have sufficient points at regions where the baseline curve shows rapid changes. Selection of a small number of basis functions is then obtained by using non-smooth at zero penalties which eliminate basis functions when they are non necessary and mainly retain basis functions whose support covers regions with sharp features.

As a fitting criterion, we will not use the standard quadratic loss which yields a regression function given by a conditional mean. To estimate the baseline in MS spectra, it is more natural to have an estimate that satisfies the property that a proportion τ of the conditional distribution of Y with respect to regressors is above the estimate. Following [21], define the following loss function as

$$\rho_\tau(u) = \tau|u|I\{u > 0\} + (1-\tau)|u|I\{u \leq 0\},$$

where $I\{\cdot\}$ is an indicator function. This loss function highlights the basic difference between the conditional mean and the conditional quantile function. Here $\tau \in (0,1)$ indicates the quantile of interest. For $\tau = 0.5$ minimization of such a loss function yields an estimate which is the median. In our context, most of the signal of interest in a spectrum lies above the baseline which is assumed to be slowly varying, and therefore it seems natural to estimate the baseline by using quantile regression with a small value of τ. While there is no criterion for establishing when most of the data lie above the baseline, a cutoff of $\tau = 0.001$ works well.

Our quantile spline estimator of B is then finally given as the minimizer of

$$\min_{\beta \in \mathbb{R}^K} \sum_{i=1}^{n} \rho_\tau (Y(x_i) - H(x_i)\beta) + \lambda \sum_{j=1}^{K} |\beta_j|.$$

Like other nonparametric smoothing methods, the smoothing parameter λ plays a crucial role on determining the trade-off between the fidelity to the data and the penalty. Two commonly criteria for choosing λ are the Schwarz information criterion [22] (SIC) and the generalized approximate cross-validation criterion [39] (GACV). In our practical implementation we have used SIC.

2.3. Spectra normalization

To remove the normalization factor N in model (1), we simply divide each denoised and baseline corrected spectra by its area under the curve (AUC) which is a standard normalizing choice in MS spectra analysis. In Figure 2, we give an example of spectra pre-processing using denoising by wavelet thresholding and baseline correction with the SIC criterion. Note that the method is fully automatic and that the peaks of the spectra are well preserved after baseline correction.

3. Multiscale detection and alignment of peaks

Spectra alignment consists in finding, for each observed spectrum, a warping function in order to synchronize all spectra before applying any other statistical inferential procedure. In this section, we present an automatic method for detecting and aligning peaks from denoised, baseline corrected and normalized spectra. This new peak detection algorithm for MS spectra analysis is based on the method developed by [4, 5], and the problem of aligning multiple spectra is formulated as a non-rigid curve registration problem using recent techniques developed in [7, 6, 8, 9, 23] .

Before introducing our framework, we briefly review few methods that have been recently proposed for addressing the alignment problem for MS spectra (see [3] for a more detailed discussion). Wavelets to represent the MS data in a multiscale framework are used in [29]. Within their framework, using a specific peak detection procedure, they first align peaks at a dominant coarser scale from multiple samples and then align the remaining peaks at a finer scale. However one may question if representing peaks at multiple scales is biologically reasonable, i.e., if peaks at coarse scales really correspond to true peptides. A similar in spirit procedure has been also developed by [12]. In [19] it is assumed that the peak variation is less than the typical distance between peaks and use a closest point matching method in peak alignment. The applicability of their method is limited by the data quality and it cannot handle large peak variation or false positive peak detection results. A recent method is the nonparametric warping model with spline functions to align MS spectra proposed recently in [18]. While the idea of using smoothing splines to model the warping function is interesting it is unclear if a smooth function with second order regularities is precise enough to describe the nonlinear shift of MS peaks encountered in practice. In [34] the authors applied a hierarchical clustering method to construct a dendrogram of all peaks from multiple samples. They cut off the dendrogram using a predefined parameter and clustered the remaining branches into different groups. They then consider the centers of

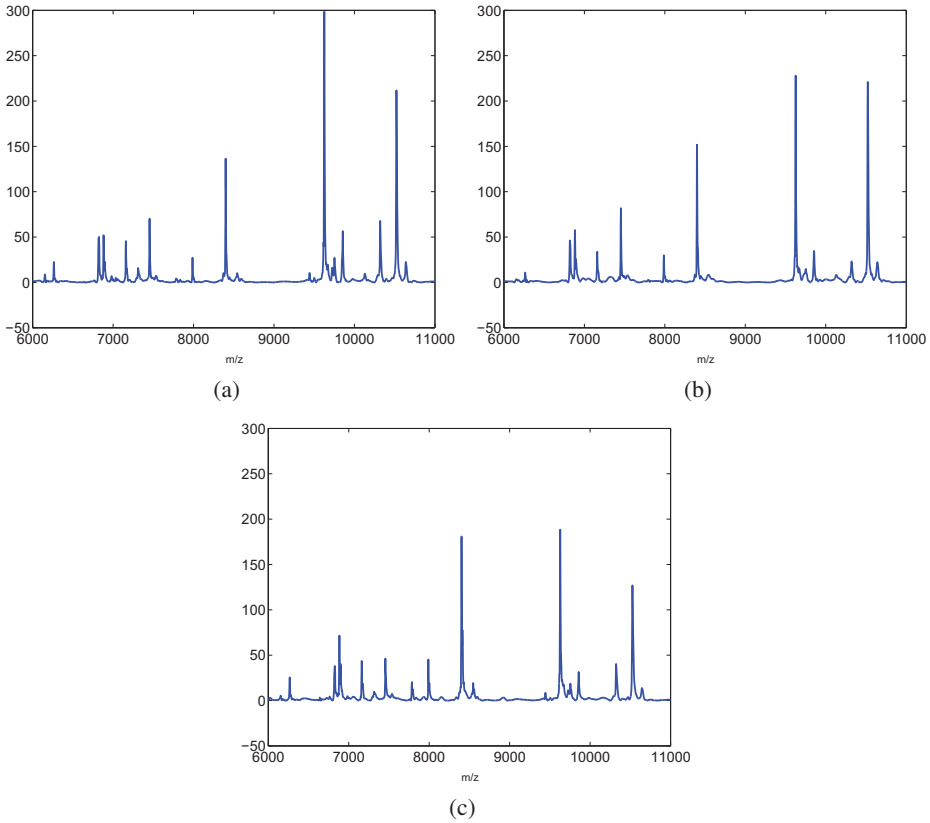

FIGURE 2. *An example of wavelet-based denoising followed by baseline correction using penalized regression quantile splines - Compare with Figure 1*

these groups as common peaks and align every peak set with respect to the common peaks. The implicit assumption behind their approach is that around each of the common peaks, the observed peaks from multiple samples obey a certain kind of distribution with the mean equal to the location of the common peak. The assumption agrees well with the motivation of peak alignment. However, the cut-off parameter and the final clustering results can be influenced by changing a few nodes in the dendrogram, while some noisy points or outliers (e.g., caused by false positive peak detection results) often cause such changes, as it is shown in the paper by [37]. To our knowledge, frameworks that are closest to the one we propose in this work are the recent approaches proposed by [33] and [38]. Both use a robust point matching algorithm to solve the alignment problem but an implicit assumption in [33] is that there exists a one-to-one correspondence among peaks in multiple spectra while [38] use a super set method to calibrate the alignment. To end this subsection let us mention also that [2] also discusses alignment methods that address proteomic data, yet provides few details and no software for their implementation.

3.1. Multiscale detection of peaks

Let us now describe, the multiscale approach proposed in [4, 5] to estimate the significant peaks of a noisy signal. This method is based on the estimation of the significant zero-crossings of the continuous wavelet transform of a noisy signal, and on a new tool, the structural intensity, proposed in [4] to represent the landmarks of a signal via a probability density function. The main modes of the structural intensity correspond to the significant landmarks of the unknown signal. In a sense, the structural intensity can be viewed as a smoothing method which highlights the significant features of a signal observed with noise.

Let f be a real-valued function, θ a smooth function with a fast decay such that $\int_{-\infty}^{\infty}\theta(u)du \neq 0$ and $\psi(x) = \frac{\partial}{\partial x}\theta(x)$ a wavelet with one vanishing moment (in our numerical experiments θ is a Gaussian density). Then, by definition, the continuous wavelet transform of f at a given scale $s > 0$ is:

$$W_s(f)(x) = \int_{-\infty}^{+\infty} f(u)\psi_s(u-x)du \text{ for } x \in \mathbb{R},$$

where $\psi_s(u) = \frac{1}{\sqrt{s}}\psi(\frac{u}{s})$. The term *zero-crossings* is used to describe any point (z_0, s_0) in the time-scale space such that $z \mapsto W_{s_0}(f)(z)$ has exactly one zero at $z = z_0$ in a neighborhood of z_0. We will call *zero-crossings line* any connected curve $z(s)$ in the time-scale plane (x, s) along which all points are zero-crossings. Now, note that if f is differentiable in an interval $[a, b]$, then for all $x \in]a, b[$

$$\lim_{s\to 0}\frac{W_s(f)(x)}{s^{3/2}} = C\frac{\partial}{\partial x}f(x), \text{ where } C = \int_{-\infty}^{+\infty}\theta(u)du \neq 0. \tag{2}$$

Hence, equation (2) shows that at small scales the zero-crossings of $W_s(f)(x)$ converge to the zeros of the derivative of f in $]a, b[$ (if any). Thus, locations of the extrema of f can be found by following the propagation at small scales of the curves $z(s)$ (see [4] for further references).

The significant peaks of denoised and baseline corrected MS spectra can be detected by estimating the significant zero-crossings that correspond to local maxima. A simple hypothesis testing procedure has been developed in [5] to estimate the zero-crossings lines of a noisy signal

at various scales. The test proposed is based on the choice of an appropriate smoothing parameter to separate, in a denoised and baseline corrected spectra, the zero-crossings of the true signal of interest $S(m/z)$ from those that are due to small fluctuations. An example of zero-crossing estimation from MS spectra is displayed in Figure 3(d,e,f).

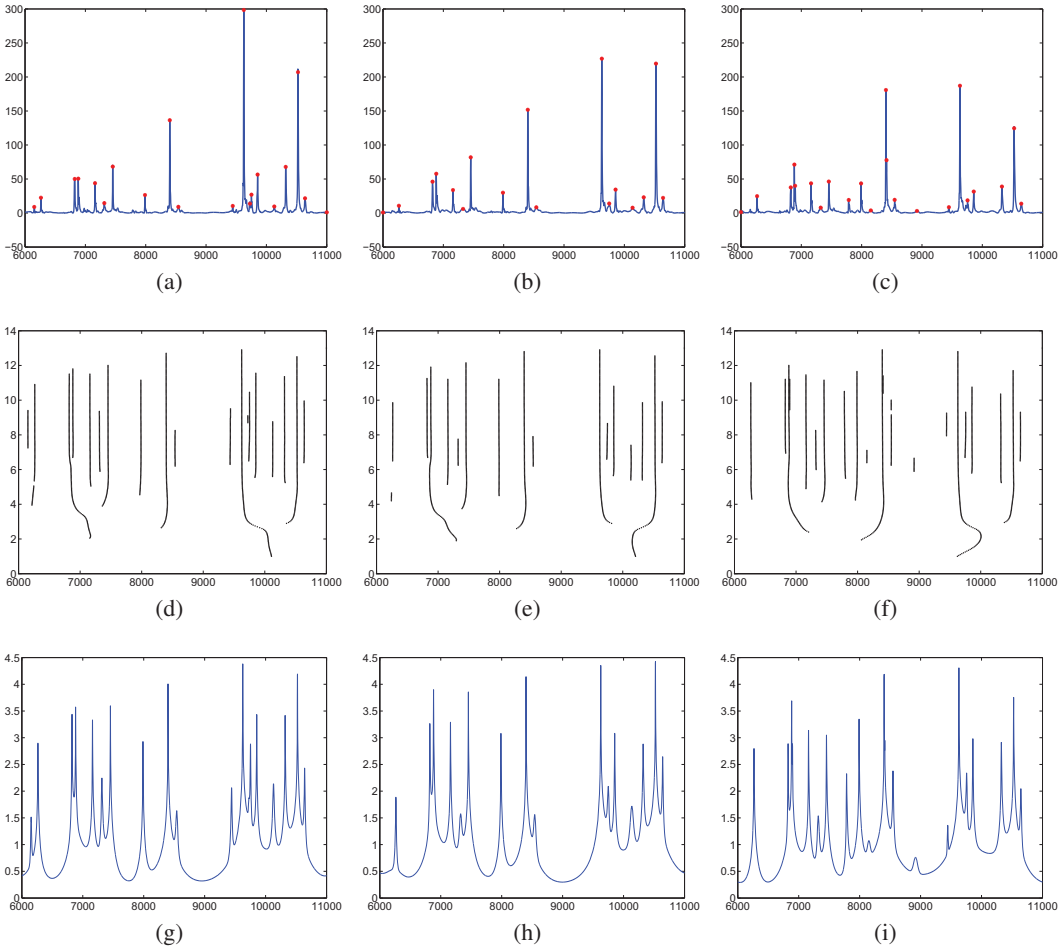

FIGURE 3. *An example of multiscale peaks detection. First row: denoised and baseline corrected spectra with red star indicating the significant peaks. Second row: estimation of significant zero-crossing at various scales (the vertical axis corresponds to the negative logarithmic scale). Third row: structural intensity of the zero-crossing (note that maxima of the structural intensities are located at the significant peaks of the MS spectra).*

However, such a procedure only gives a visual representation that indicates "where" the peaks are located, but there is generally no analytical expression of the zero-crossings lines in a closed form. The structural intensity is a powerful tool introduced in [4] to identify the limits of the zero-crossing lines when they propagate to small scales. The structural intensity method consists in using the zero-crossing of a signal at various scales to compute a "density" whose local maxima will be located at the peaks of f.

More precisely, the structural intensity is defined to be:

$$G_z(x) = \sum_{i=1}^{\hat{p}} \int_{\hat{s}_i^1}^{\hat{s}_i^2} \frac{1}{s} \theta\left(\frac{x - \hat{z}_i(s)}{s}\right) ds, \tag{3}$$

where $\hat{p}$ is the number of estimated zero-crossings lines $\hat{z}_i(.), i = 1, \ldots, \hat{p}$ and $[\hat{s}_i^1, \hat{s}_i^2]$ represent the supports of these lines in the time-scale plane. It can be seen from Figure 3 that, for small values of the scale s, the zero-crossings lines $\hat{z}_i(s), i = 1, \ldots, \hat{p}$ are close to the locations of the peaks of a signal. Therefore, for small values of s, the $\hat{z}_i(s)$'s can be viewed as random variables which are all located near the peaks of a signal. The above definition (3) of structural intensity can thus be interpreted as a kind of kernel density estimator with scale s playing the role of a varying bandwidth. Since the $\hat{z}_i(s)$'s are concentrated around the same locations for small s, the main modes of such a density correspond to the limits of the $\hat{z}_i(s)$'s for small values of s. The significant peaks of the spectra f are then defined as the local maxima of $G_z(x)$ on $[0,1]$. The structural intensity is therefore a tool to identify the limits of the lines $\hat{z}_i(.), i = 1, \ldots, \hat{p}$ in the time-scale plane, Figure 3(d,e,f). In Figure 3, one can see that the local maxima of the structural intensity correspond to the significant peaks of the spectra.

3.2. Spectra alignment and peaks correspondence

To identify differentially expressed proteins in samples of diseased and healthy individuals, biologically significant comparisons and conclusions are all based on the alignment results of spectra. However, MS spectra alignment is difficult even after instrument calibration with internal markers because the mass errors vary with m/z in a nonlinear fashion as a result of experimental and instrumental complexity and data variation. As displayed in Figure 4(a), the three spectra used to illustrate our methodology clearly have significant peaks that are not aligned which makes difficult the identification of common peaks to compare multiple spectra. In Figure 4(b), we display the result of the alignment procedure that we describe hereafter.

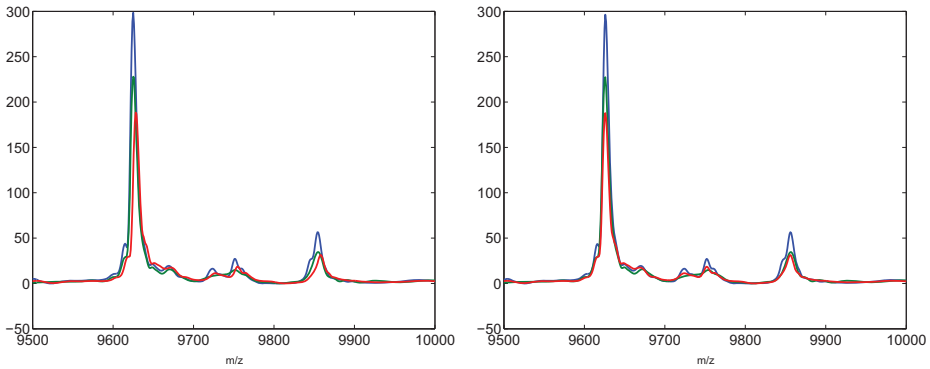

FIGURE 4. *(a) An example of mis-alignment of significant peaks. (b) Aligned spectra using warping functions minimizing the criterion (4). For better visualization we only plotted a small part of the spectra*

Recall that spectra alignment consists in finding, for each observed spectrum, a warping function in order to synchronize all spectra before applying any other statistical inferential procedure (a warping function is a strictly monotone transformation of the interval onto which the spectra are defined). Such an issue is usually referred to as the curve registration problem in statistics and various methods have been proposed, see e.g. [20], [16], [28], [5] and references therein. Classically, to align more than two curves, one uses the average of the observed curves as a reference template onto which all observed are registered. We have used this approach in [3] combined with the automatic landmark-based registration technique proposed in [5]. However, in this paper, we suggest a new method to align multiple spectra that does not require the use of a reference template. For this, we will use a parametric representation of warping functions proposed in [8, 7] combined with a new criterion proposed in [6, 9, 23] to align a set of signals that does not require the use of a reference template.

Let us first describe the parametric model for warping functions proposed in [8, 7]. Let $\Omega \subset \mathbb{R}$ be a compact interval representing the support of the spectra and denote by x the mass to charge ratio variable m/z. Let $v(\cdot)$ be a smooth parametric function given by a linear combination of known basis functions $\{h_j : \Omega \to \mathbb{R},\ j = 1,\ldots,K\}$,

$$v(x) = \sum_{j=1}^{K} a_j h_j(x).$$

The function v is thus parametrized by the set of coefficients $\mathbf{a} = (a_1,\ldots,a_K)^T \in \mathbb{R}^K$ and we write $v = v_{\mathbf{a}}$ to stress this dependency. In what follows, it will be assumed that the basis functions are continuously differentiable on Ω and such that h_j and h_j' vanish at the boundaries of Ω. For the h_j's we took in our simulations a set of $K = 20$ B-spline functions that form a basis for polynomial splines of degree $p = 3$ with equally-spaced knots $\tilde{x}_1 < \cdots < \tilde{x}_q$ on Ω with $q = 16$. Then, let $x \in \Omega$ and for $t \in [0,1]$ consider the following ordinary differential equation (ODE)

$$\frac{\partial}{\partial t}\phi(t,x) = v_{\mathbf{a}}(\phi(t,x))$$

with initial condition $\phi(0,x) = x$. Then, it can be shown (see [8, 7]) that for any $t \in [0,1]$ the solution of the above ODE is unique and such that $x \mapsto \phi(t,x)$ is a strictly increasing function on Ω such that $\phi(t,\Omega) = \Omega$. Then, denote by $\phi_{\mathbf{a}}(x) = \phi(1,x)$ the solution at $t = 1$. In this way we thus obtain a warping function $\phi_{\mathbf{a}}$ that is parametrized by the set of coefficients $\mathbf{a} \in \mathbb{R}^K$.

Now, to compute warping functions to synchronize the spectra, we will search to align the structural intensities displayed in Figure 3 rather than the spectra themselves. As the structural intensities have their local maxima located at the significant peaks of the spectra this will force the search of warping functions that put into correspondence the main peaks of the spectra. Assume that one has to align a set of N spectra with corresponding structural intensities $G_1,\ldots,G_N$. Then N warping functions $\phi_1,\ldots,\phi_N$ can be computed by minimizing the following cost that has been recently proposed in [6, 9, 23] as a new matching criteria for curve and image warping:

$$(\hat{\mathbf{a}}_1,\ldots,\hat{\mathbf{a}}_N) = \underset{(\mathbf{a}_1,\ldots,\mathbf{a}_N)\in\mathbb{R}^{K\times N}}{\arg\min} \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{n}\left(G_i \circ \phi_{\mathbf{a}_i}(x_k) - \frac{1}{N}\sum_{i'=1}^{N} G_{i'} \circ \phi_{\mathbf{a}_{i'}}(x_k)\right)^2 \tag{4}$$

where $x_1, \ldots, x_n$ are the design points in Ω where the spectra are observed. Then, the warping functions are given by

$$\phi_i = \phi_{\hat{\mathbf{a}}_i} \text{ for } i = 1, \ldots, N.$$

Minimization of the criterion (4) over the set $\mathbb{R}^{K \times N}$ is possible by a gradient descent algorithm with an adaptive step as described in [9]. Note that minimizing (4) automatically yields warping functions $v_i, i = 1, \ldots, N$ and compute a reference template given by

$$\frac{1}{N} \sum_{i=1}^{N} G_i \circ \phi_i(x) \text{ for } x \in \Omega.$$

Then the warping functions $\phi_1, \ldots, \phi_N$ can be used to align the spectra. An example of such alignment is displayed in Figure 4(b). Then, to identify common peaks to a set spectra, one simply takes the significant peaks given by the structural intensities that are at the same locations after alignment using the warping functions $\phi_1, \ldots, \phi_N$. An example of common peaks detection by this procedure is displayed in Figure 5. Locations and intensities of such peaks can then be used as biomarkers and individual features for further multi-dimensional statistical analysis.

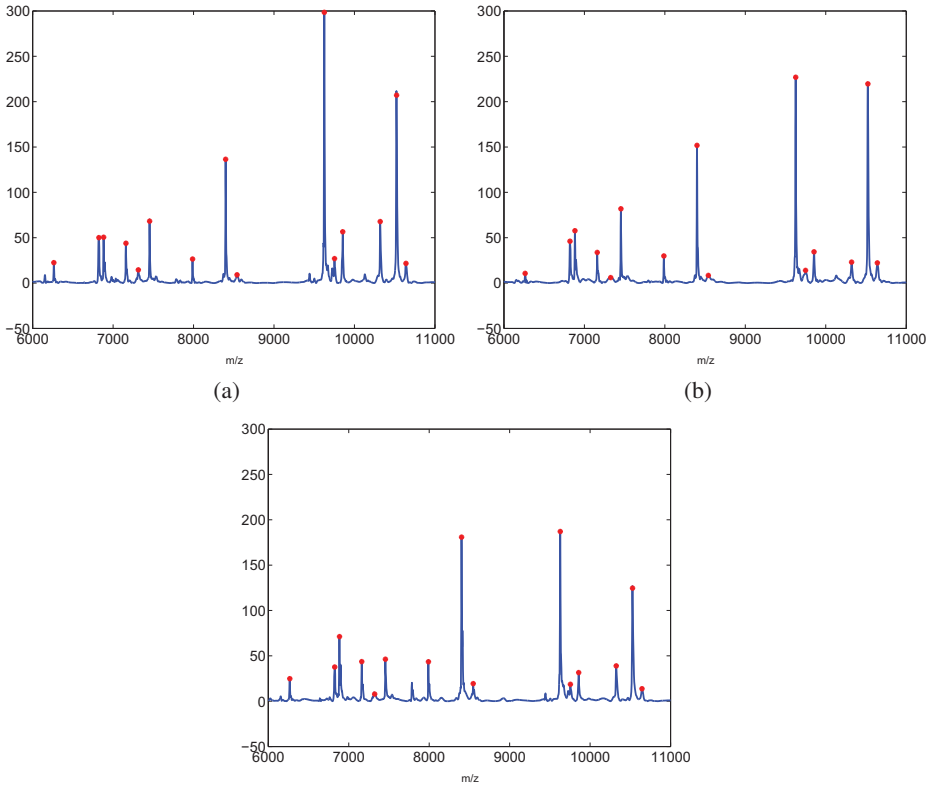

FIGURE 5. *An example of common peaks detection. The blue curves are denoised and baseline corrected spectra (not aligned) with red stars indicating the significant common peaks found using our alignment procedure.*

3.3. *Comparison with another algorithm for automatic peaks correspondence and spectra alignment*

Let us now compare our method with the algorithm proposed by Jeffries in [18] which also performs automatic peaks correspondence and spectra alignment. For this purpose, we propose to test the two methods on the data described in [18] that are available from:
`http://krisa.ninds.nih.gov/alignment/`

Let us first describe these data. As part of a larger study examining proteomic spectra from healthy individuals and those with multiple sclerosis, reference samples from a large pool of serum were included as part of a quality control procedure. As patient and control samples were processed, a few spectra were consistently drawn from this common, fixed reference pool and analyzed to alert investigators to deviations related to sample processing. Ideally, all the spectra from the reference samples should look very similar. Samples were processed on six distinct days using identical calibration procedures, staff, equipment, and sample handling techniques. Samples from the first four days were processed within a single week while samples for the last two days were processed approximately 2 and 3 months later (see [18] for further details). We have chosen to consider the simple problem of aligning two spectra: a reference spectrum from the third day and a spectrum for the fifth day. Both spectra are displayed in Figure 6. It can be seen that the data produced on the fifth day are not well aligned with those of the third.

Then, for sake of completeness, let us briefly recall the method proposed by Jeffries in [18] for peaks correspondence and spectra alignment. This approach uses a reference spectrum (e.g. data from the third day) to which the other spectrum (data from the fifth day) will be compared and aligned. To begin with, define a subrange of the mass values say $[5000, 5000+a]$ Daltons (with $a = 500$ or $a = 250$). For this range locate the largest peak in the spectrum to be aligned and note its location denoted as p_1. Then, consider all the peaks in the reference spectrum that are located within a fixed window around p_1, say $[0.98p_1, 1.02p_1]$. If there are k peaks within this 2% window they are denoted as $m_1^1, \ldots, m_k^1$. For each of these peaks consider a small window centered at m_j^1, such as $[0.95m_j^1, 1.05m_j^1]$, for $j = 1, \ldots, k$. For each one of these windows, compute the correlation coefficients of the intensities in the reference spectrum over this mass range with the intensities over a small window centered about p_1, say $[0.95p_1, 1.05p_1]$, in the spectrum to be aligned. From these k correlation coefficients choose the corresponding target peak as that with the highest correlation. This procedure is then repeated on the ranges $[5000+a(u-1), 5000+au]$ for $u = 2, \ldots, q$ with $q = 10$ for $a = 500$ and $q = 20$ for $a = 250$. Thus, in this way, one obtains a set of q peaks locations $p_1, \ldots, p_q$ in the spectrum to be aligned, and a set of q peaks locations $m_1, \ldots, m_q$ in the reference spectrum. Further details regarding implementation are available at http://krisa.ninds.nih.gov/alignment/.

Then, a warping function $\hat{\phi}$ to align the spectra from the fifth day onto the reference spectra (from the third day) can be computed by finding the mapping $\hat{\phi}$ which minimizes the following cost:

$$\hat{\phi} = \arg\min_{\phi \in \Phi} \sum_{u=1}^{q} (p_u - \phi(m_u))^2,$$

where Φ denotes a set of smooth increasing functions (e.g. monotone cubic splines or homeomorphic splines, see [7] for further details on this issue).

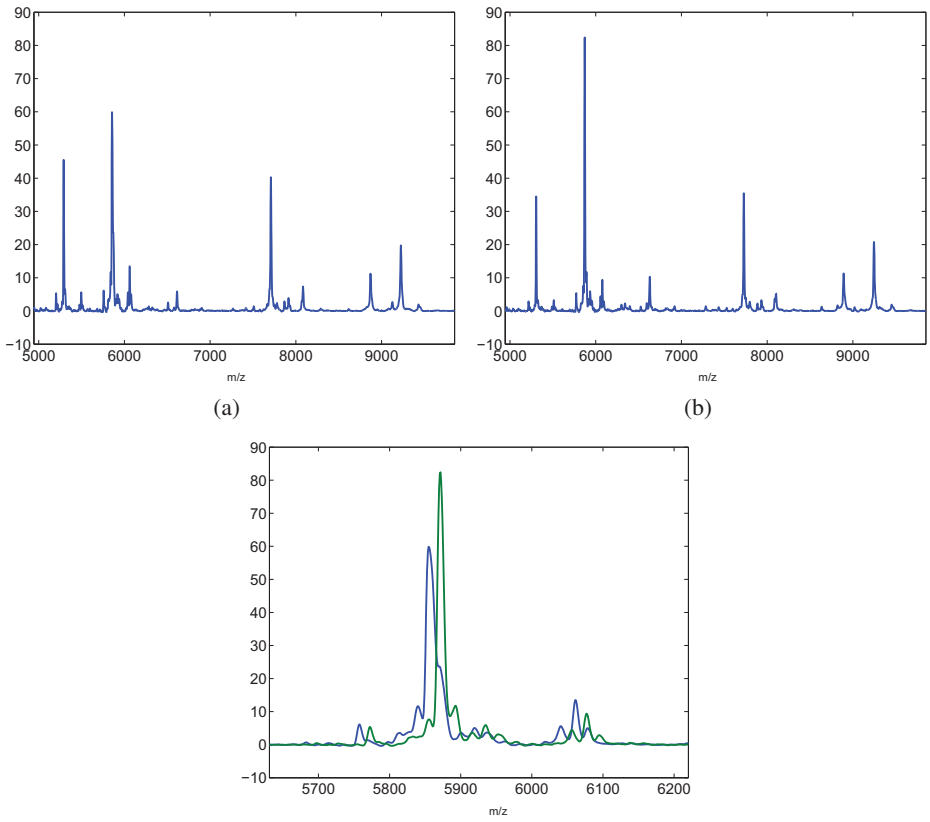

FIGURE 6. *SELDI-TOF proteomic data from [18]: (a) reference spectrum from the third day, (b) spectrum from the fifth day, (c) superposition of the two spectra on the interval* $[5700, 6200]$ *(mass to charge ratio* (m/z) *in Daltons).*

The results obtained using our method and the above described procedure with $q = 10$ and $q = 20$ are displayed in Figure 7, Figure 8 and Figure 9 respectively. It can be seen that both methods perform very well for the alignment of the two spectra. However, when comparing the common peaks found by our method and those found by Jeffries' procedure, it can be seen that the results are very much different. Our method detects all the significant common peaks that would have been found manually by a visual comparison of the spectra. This is not the case for Jeffries' procedure: for either $q = 10$ or $q = 20$, a significant number of spurious common peaks are found in intervals without visually significant peaks, and the method also fails in detecting truly significant common peaks. This example shows that our method clearly outperforms Jeffries' procedure, in the sense that it not only performs spectra alignment but also allows an automatic correspondence between truly significant peaks.

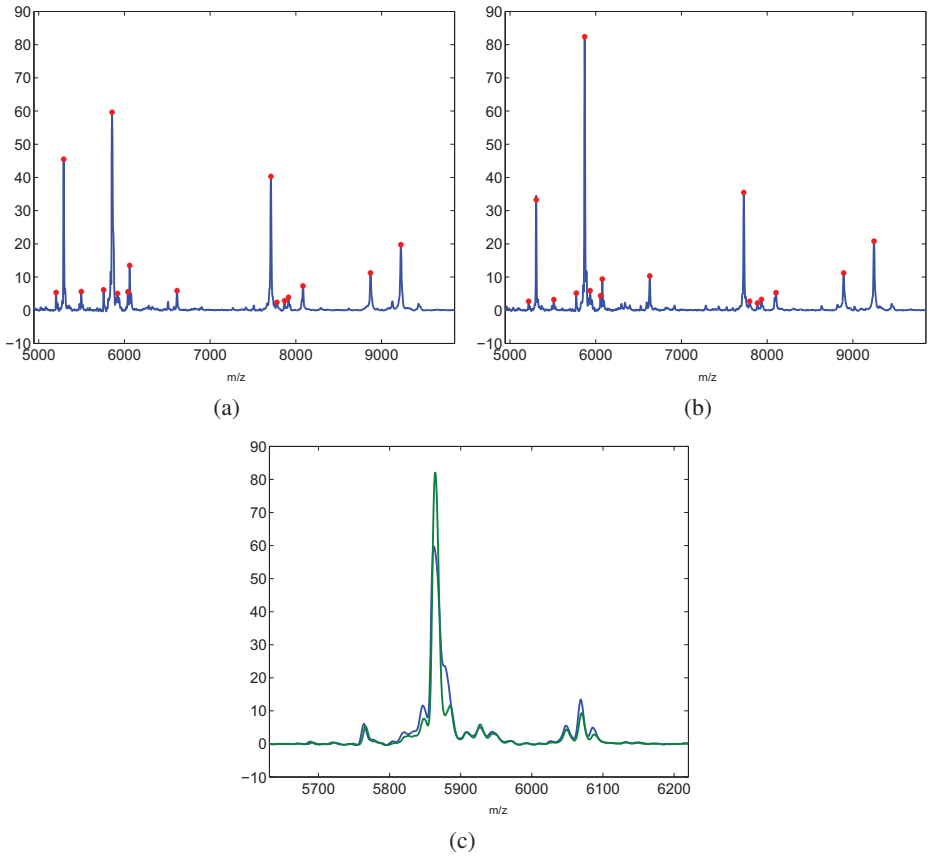

FIGURE 7. *Results of spectra alignment and peaks correspondence (with red stars indicating the common peaks) using our method: (a) reference spectrum from the third day, (b) spectrum from the fifth day, (c) alignment of the two spectra on the interval* $[5700, 6200]$ *(Mass/Charge Weight in Daltons).*

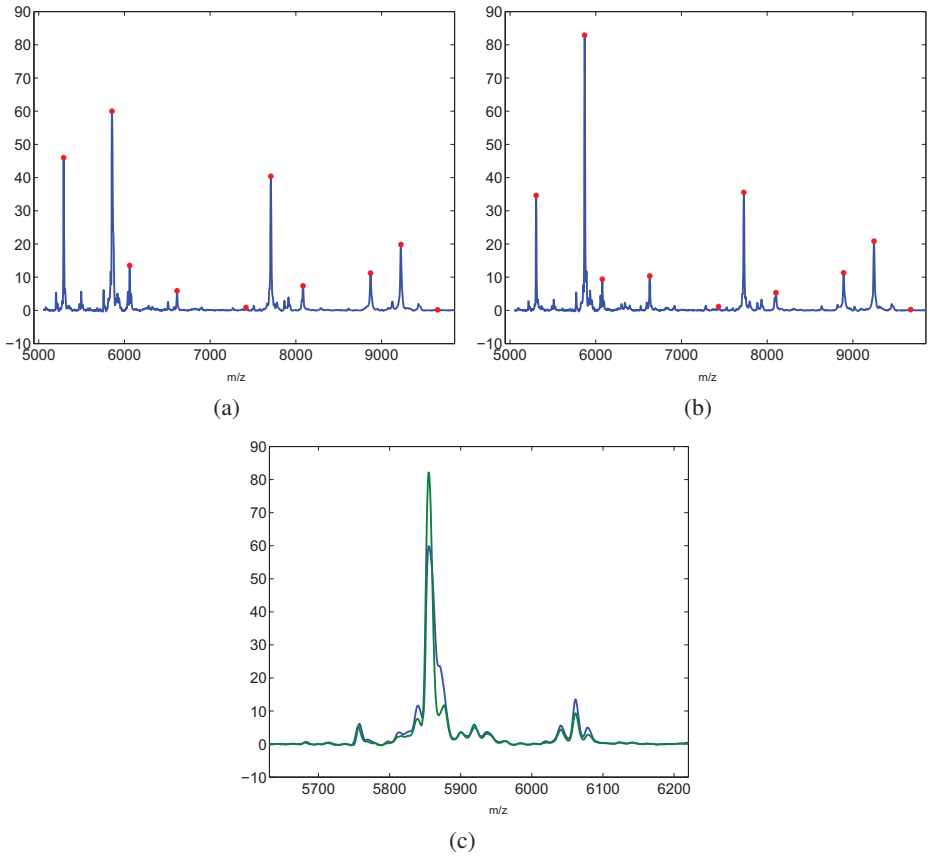

FIGURE 8. *Results of spectra alignment and peaks correspondence (with red stars indicating the common peaks) using Jeffries' method with* $q = 10$*: (a) reference spectrum from the third day, (b) spectrum from the fifth day, (c) alignment of the two spectra on the interval* $[5700, 6200]$ *(Mass/Charge Weight in Daltons).*

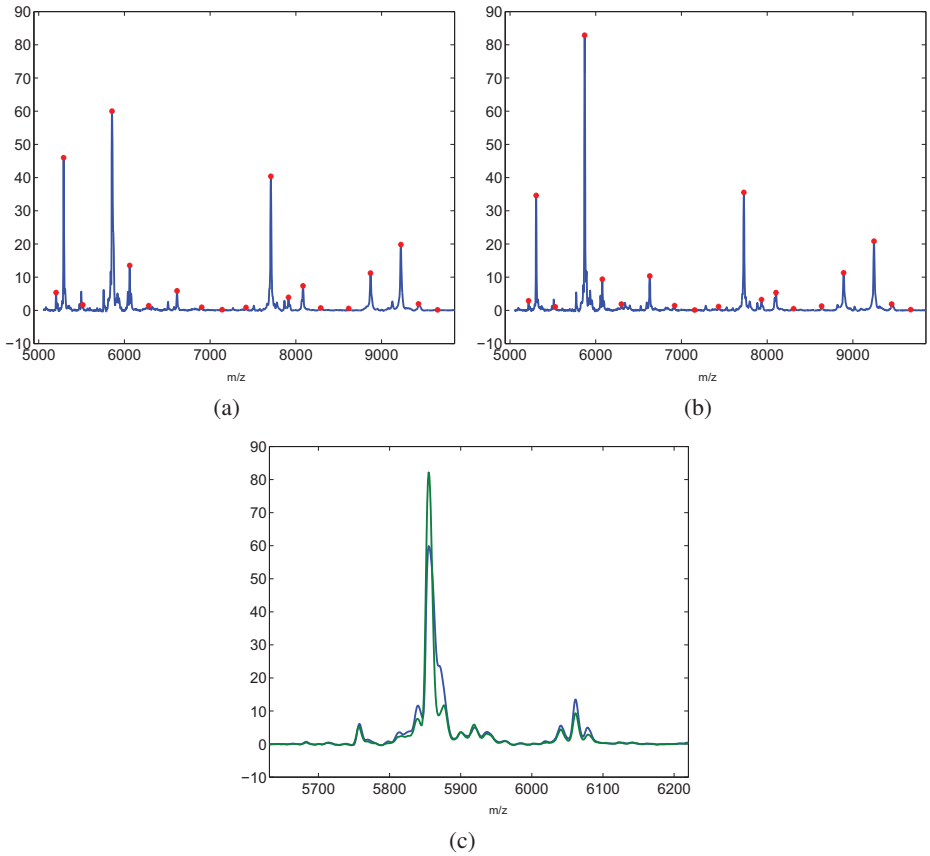

FIGURE 9. *Results of spectra alignment and peaks correspondence (with red stars indicating the common peaks) using Jeffries' method with* $q = 20$*: (a) reference spectrum from the third day, (b) spectrum from the fifth day, (c) alignment of the two spectra on the interval* $[5700, 6200]$ *(Mass/Charge Weight in Daltons).*

4. Conclusion

In this paper we have first reviewed some commonly used statistical methods for pre-processing of raw MS data. We have also proposed to use a multi-scale detection procedure based on the continuous wavelet transform to improve peak detection in observed high-resolution spectra. The ability of this multi-scale detection procedure to detect efficiently potential peaks has allowed us to investigate a new method for aligning multiple spectra without the need of a reference template. We applied this method to several spectra with baseline distortion, and demonstrate it is effective for spectra with over- crowded peaks. Our study extends the application scope of both the continuous wavelet transform and the deformation based warping methodology for spectra alignment. We have also compared our method with another procedure for spectra alignment and automatic peaks correspondence. Both methods perform similarly for the alignment of two spectra, but the results indicate that our method is clearly much better for the identification of truly significant common peaks. This example shows some of the benefits of our approach over existing methods in the literature. We hope that the methods presented here will stimulate further investigation into the development of better procedures for multi-scale modeling and analysis of high-resolution MS spectra.

Acknowledgements

This work was supported in part by a Grant of the Interuniversity Attraction Pole IAP P6/03 of the Belgium Government. We would like also to thank Philippe Besse for fruitful discussions on statistical methods for the analysis of high-resolution MS spectra. J. Bigot would like to thank the Center for Mathematical Modeling and the CNRS for financial support and excellent hospitality while visiting Santiago where part of this work was carried out. The authors would like to thank the both referees for their relevant comments.

References

[1] H. Abbink Spaink, T.T. Lub, R.P. Otjes, and H.C. Smith. Baseline correction for second-harmonic detection with funable diode lasers. *Anal. Chim. Acta*, 183:141–151, 1986.

[2] Sauve. A.C. and T.P. Speed. Normalization, baseline correction and alignment of high-throughput mass spectrometry data. In *Proceedings Gensips*, 2004. in press.

[3] A. Antoniadis, J. Bigot, S. Lambert-Lacroix, and F. Letué. Nonparametric pre-processing methods and inference tools for analyzing time-of-flight mass spectrometry data. *Current Analytical Chemistry*, 3:127–147, 2007.

[4] J. Bigot. A scale-space approach with wavelets to singularity estimation. *ESAIM: PS*, 9:143–164, 2005.

[5] J. Bigot. Landmark-based registration of curves via the continuous wavelet transform. *Journal of Computational and Graphical Statistics*, 15(3):542–564, 2006.

[6] J. Bigot and S. Gadat. A deconvolution approach to estimation of a common shape in a shifted curves model. *Annals of Statistics*, to be published, 2010.

[7] J. Bigot and S. Gadat. Smoothing under diffeomorphic constraints with homeomorphic splines. *SIAM Journal on Numerical Analysis*, 48:224–243, 2010.

[8] J. Bigot, S. Gadat, and J.M. Loubes. Statistical M-estimation and consistency in large deformable models for image warping. *Journal of Mathematical Imaging and Vision*, 34:270–290, 2009.

[9] J. Bigot, F. Gamboa, and M. Vimond. Estimation of translation, rotation and scaling between noisy images using the fourier mellin transform. *SIAM Journal on Imaging Sciences*, 2:614–645, 2009.

[10] F.T. Chao and A.K.M. Leung. *Application of wavelet transform in processing chromatographic data*. Walczak B (ed.) Wavelets in Chemistry, Elsevier Science, 2000.

[11] R.R. Coifman and D.L. Donoho. Translation invariant de-noising. *Lecture Notes in Statistics*, 103:125–150, 1995.

[12] K.R. Coombes, S. Tsavachidis, J.S. Morris, K. A. Baggerly, and R. Kobayashi. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra using the undecimated discrete wavelet transform. *Proteomics*, 41:4107–4117, 2005.

[13] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[14] C. De Boor. *A pratical guide to splines*. Vol. 27 of Applied Mathematical Sciences, Springer-Verlag, New-York, 1978.

[15] P. Dierckx. *Curve and surface fitting with splines*. Clarendon, Oxford, 1993.

[16] T. Gasser and A. Kneip. Searching for structure in curve samples. *Journal of the American Statistical Association*, 90(432):1179–1188, 1995.

[17] C. Heipke. Overview of image matching techniques. In *OEEPE - Applications of Digital Photogrammetric Workstations, Proceedings, Lausanne, Switzerland*, pages 173–191, 1996.

[18] N. Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21(14):3066–3073, 2005.

[19] K.J. Johnson, B.W. Wright, K.H. Jarman, and R.E. Synovec. High-speed peak matching algorithm for retention time alignment of gas chromatographic data. *Journal of Chromatography A*, 996:141–155, 2003.

[20] A. Kneip and T. Gasser. Statistical tools to analyze data representing a sample of curves. *Annals of Statistics*, 20(3):1266–1305, 1992.

[21] R. Koenker and G. Basset. Regression quantiles. *Econometrica*, 1:33–50, 1978.

[22] R. Koenker, P. Ng, and S. Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994.

[23] J.M. Loubes, E. Maza, and F. Gamboa. Semi-parametric estimation of shifts. *Electronic Journal of Statistics*, 1:616–640, 2007.

[24] X.G. Ma and Z.X. Zhang. Application of wavelet transform to background correction in inductively coupled plasma atomic emission spectrometry. *Anal. Chim. Acta*, 485(2):233–239, 2003.

[25] S.G. Mallat. *A Wavelet Tour of Signal Processing. 2nd ed.* San Diego: Academic Press, 1999.

[26] J. Padayachee, V. Prozesky, W. von der Linden, M.S. Nkwinika, and V. Dose. Bayesian PIXE background subtraction. *Nucl. Instrum. Methods Phys. Res. B*, 150:129–135, 1999.

[27] Y. Qu, B.L. Adam, M. Thornquist, J.D. Potter, M.L. Thompson, Y. Yasui, J. Davis, P.F. Schellhammer, L. Cazares, M. Clements, G.L.Jr Wright, and Feng Z. Multiscale processing of mass spectrometry data. *Biometrics*, 59:143–151, 2003.

[28] J.O. Ramsay and X. Li. Curve registration. *Journal of the Royal Statistical Society, Series B*, 60:351–363, 1998.

[29] T. W. Randolph and Y. Yasui. Multiscale processing of mass spectrometry data. *Biometrics*, 62(2):589–597, 2006. in press.

[30] A. Rouh, M.A. Delsuc, G. Bertrand, and J.Y. Lallemand. The use of classification in baseline correction of ft nmr spectra. *J. Magn. Reson. Ser. A*, 102:357–359, 1993.

[31] A.F. Ruckstuhl, M.P. Jacobson, R.W. Field, and J.A. Dodd. Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 68(2):179–193, 2001.

[32] S. Sardy, D. B. Percival, A.G. Bruce, H.Y. Gao, and W. Stuetzle. Wavelet denoising for unequally spaced data. *Statistics and Computing*, 9(1):65–75, 1999.

[33] B. Saussen, M. Kirchner, H. Steen, J.A. Jebanathirajah, and F.A. Hamprecht. The rpm package: aligning LC/MS mass spectra with R. In *Interdisciplinary Center for Scientific Computing, University of Heidelberg, Germany UseR2006*, 2006.

[34] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q. Le. Sample classification from protein mass spectrometry, by peak probability contrasts. *Bioinformatics*, 20(17):3034–3044, 2004.

[35] E.H. van Veen and M.T.C. de Loos-Vollebregt. Application of mathematical procedures to background correction and multivariate analysis in inductively coupled plasma-optical emission spectrometry. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 53(5):639–669, 1998.

[36] W. W. Dietrich, C.H. Rüdel, and M. Neumann. Fast and precise automatic baseline correction of one- and two-dimensional NMR spectra. *J. Magn. Reson.*, 91:1–11, 1991.

[37] W. Yu, B. Wu, N. Lin, K. Stone, K. Williams, and H. Zhao. Detecting and aligning peaks in analyzing maldi mass spectrometry data. *Computational Biology and Chemistry*, 30:27–38, 2006.

[38] W. Yu and H. Zhao. Aligning spectral peaks in mass spectrometry data with a robust point matching approach. In *In 52nd ASMS Conference on Mass Spectrometry and Allied Topics, Nashville, TN, May*, pages 23–27, 2004.

[39] M. Yuan. Gacv for quantile smoothing splines. *Computational Statistics and Data Analysis*, 50(3):813–829, 2006.