Align – A Software Tool For Mass Spectra Preprocessing

Michal Marczyk, *Institute of Automatic Control, Silesian University of Technology*

Abstract

An important issue in analyzing mass spectra is to extract as much information as possible from a limited number of samples, which consist usually unknown amount of error, so appropriate use of bioinformatics tools becomes very critical. Here we present a graphical tool created in a Matlab 7.7 environment. It is a fully equipped application called Align, which is ready for preprocessing of different types of mass spectra. It gives an opportunity to correct a baseline, remove the noise from the data, normalize all samples and align the spectra. User can watch a heatmap, all spectra together or a mean spectrum in every phase of preprocessing. There is also a possibility to check the differences of intensities of individual spectra between the reference spectrum. Final result of the mass spectra preprocessing are peaks found in the mean spectrum.

Some of the alignment algorithms were given from the authors, some were written based on available documentation. Functions from Bioinformatics toolbox were used for a baseline correction and smoothing.

1. Introduction

Mass spectrometry (MS) has increasingly become the method of choice for analysis of complex protein samples. Its ability to identify and quantify thousands of proteins from complex samples holds special promise for the discovery of novel biomarkers. A significant disadvantage of SELDI- and MALDI-based approaches is the difficulty of moving from detection of differential pattern to the identity of the peaks comprising the pattern (Aebersold et al., 2003; Paweletz et al., 2001; Rifai et al., 2006). The detection of patterns in the multitude of mass peaks that arise from complex clinical samples depends on the uniformity of instrumental response in both mass and intensity. Hence, the instrument must be thoroughly calibrated. The fundamental weaknesses are excessive background noise, reduced signal-to-noise at high masses, misassignment of peak masses, formation of multiple chemical adducts, and substantial overlap of peaks resulting from low resolution. All of these effects make it much more difficult to distinguish and identify peaks in the mass spectrum (Malyarenko et al., 2004).

Several papers in the literature were devoted to comparing efficiency of different approaches for extraction of biomarkers (Yang et al., 2009; Meuleman et al., 2008; Cruz-Marcelo et al., 2008). However, no standard method has been established so far regarding the preprocessing steps, including the order in which the steps might be performed. Furthermore, the preprocessing of the data before peak detection may differ from the preprocessing prior to peak quantification. Because of the potential applications of mass spectrometry studies, the development of algorithms for pre-processing MS data has been an active area of research (Coombes et al., 2007; Du et al., 2006; Li et al., 2005; Malyarenko et al., 2004; Wong et al., 2005).

2. Methods

Some preprocessing steps are carried out by using the Matlab Bioinformatics toolbox functions. For a baseline correction – msbackadj (estimates the baseline within multiple shifted windows and regresses the varying baseline to the window points using a spline approximation), for smoothing – mssgolay (Savitzky and Golay, 1964). For a normalization we firstly implement only Total Ion Current method, but Cairns et al. (2008) pointed out that it can mask some biological differences. So we also include some methods, which normalizes each spectra by standard deviation of its intensities (Meuleman *et al.,* 2008).

Peak detection is performed by using mean spectrum. This procedure is commonly used and has many advantages (Morris et al. 2005). We use an algorithm (Coombes et al. 2003), which first finds peaks in the spectrum using the first derivative. Next it checks a ratio of maximum to left and right side minimum separately. Small amplitude peaks, where none of these ratios is greater than given threshold value, are deleted. We estimate the noise level across the spectrum using a median filter and compute the signal to noise ratio (SNR) at each local maximum by taking the ratio of the intensity at the maximum to the estimated local noise. All local maxima with SNR greater than given value are considered peaks. Noise reduction is obtained by replacing similar intensity

peaks, which are in close m/z neighbourhood by one, highest peak. It is also possible to delete peaks of intensity lower than a given threshold. Results of mass spectra preprocessing are shown on fig. 1.

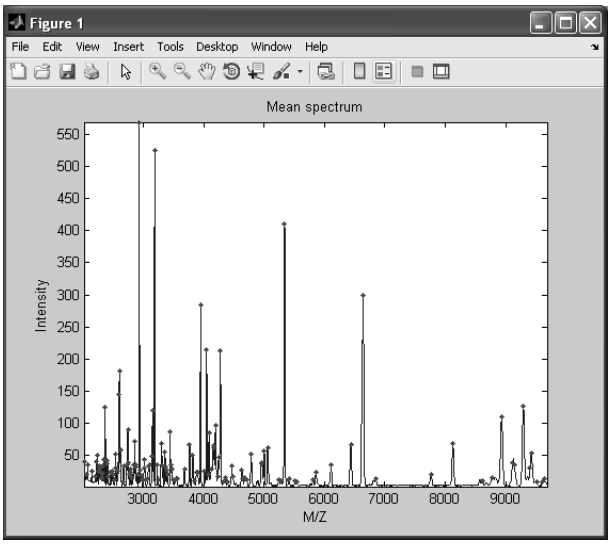

Fig.1. Mean spectrum with obtained peaks.

2.1 Alignment algorithms

From the alignment algorithms described in the literature, six are selected. Some methods are already used in spectra preprocessing software: SpecAlign (method using Fast Fourier Transform) and PrepMS (method based on using reference peaks in the spectra). Other methods are warping algorithms, which originate from chromatographic and NMR spectra analysis.

Peak Alignment by Fast Fourier Transform (PAFFT) and Recursive Alignment by Fast Fourier Transform (RAFFT) were proposed by Wong et al. (2005). Their advantage is that FFT cross-correlation is a fast and reliable method for shift estimation. The algorithm divides the spectrum into an arbitrary number of segments such that the shift in each signal can be estimated independently. The choice of minimum segment size and maximum shift allowed is not trivial. RAFFT algorithm, which makes use of a recursive segmentation model, eliminates the need to estimate the parameters.

Top five peaks alignment (TOP5) is used in application created by Karpievitch et al. (2007). It uses msalign function, which is given in Bioinformatics toolbox (Monchamp et al. 2007). In this algorithm the peaks in a mass spectrum are aligned to the reference peaks, resulting from finding top five peaks in mean spectrum.

Correlation Optimized Warping (COW) was introduced by Nielsen et al. (1998). First, both signals (target and reference) are divided into a certain number of parts. Each segment is then warped (i.e. stretched or compressed) by linear interpolation. The maximum length increase or decrease in a segment is controlled by the slack parameter. The quality of the alignment is determined by calculating the correlation coefficient between sections from both signals.

Semi-parametric Time Warping (STW) is based on a parametric time warping (Eilers, 2004). A warping function is used to the alignment of two signals. It consists of a series of B-splines, which are constructed from polynomial pieces, joined at certain values (Eilers and Marx, 1996). The aim of STW is to choose the warping coefficients, such that the sum of squared residuals between signals is minimized

The main idea of the Fuzzy Warping (FW) method relies on fuzzy sets theory. The procedure introduced by Walczak, et al. (2005) is iterative. It alternates between fuzzy matching and calculation of transform parameters. Fuzzy matching is augmented with the Sinkhorn's normalization procedure (Sinkhorn, 1964), which allows estimation of one-to-one peaks' correspondence. Similarity of corresponding peaks is then used to weigh the parameters of the global transform.

3. Program overview

Main view of the application is presented in fig. 2. First step is to load a desirable number of spectra from specified folder. After loading they are resampled to a common m/z range. User can watch all spectra with calculated sum of ions and z-score from the mean sum of ions, by pressing data button. Basic on this parameter, user can remove individual spectra from the base. There is also a possibility to find an outlier spectra, which differ in sum of ions in a specified moving window. Data binning can decrease size of the spectra by reduction of data points with averaging.

There is a fixed order of implementing the preprocessing steps. First the baseline correction is performed. All parameters of msbackadj Matlab function can be tuned by the user. Next step is noise removal. In Savitzky - Golay filter used user can change degree of the polynomial fitted to the points in the moving frame and the size of this frame. Third step is normalization using one of the six method analyzed by Meleumann, et al. (2008). Last step is a spectra alignment using one of the six methods with specified parameters. User can align spectra to the mean spectrum or load a reference spectrum. After the preprocessing user can find a peaks in a mean spectrum and view obtained results. There is an opportunity to watch an individual spectra or the mean spectrum with early obtained peaks. Different types of results representation are heatmap and differences of intensities of individual spectra between the reference spectrum. As a reference user can use mean spectrum or spectrum loaded from a file. At every stage of the preprocessing, there is a possibility to reload raw spectra, undo last preprocessing step, save the spectra and obtained peaks.

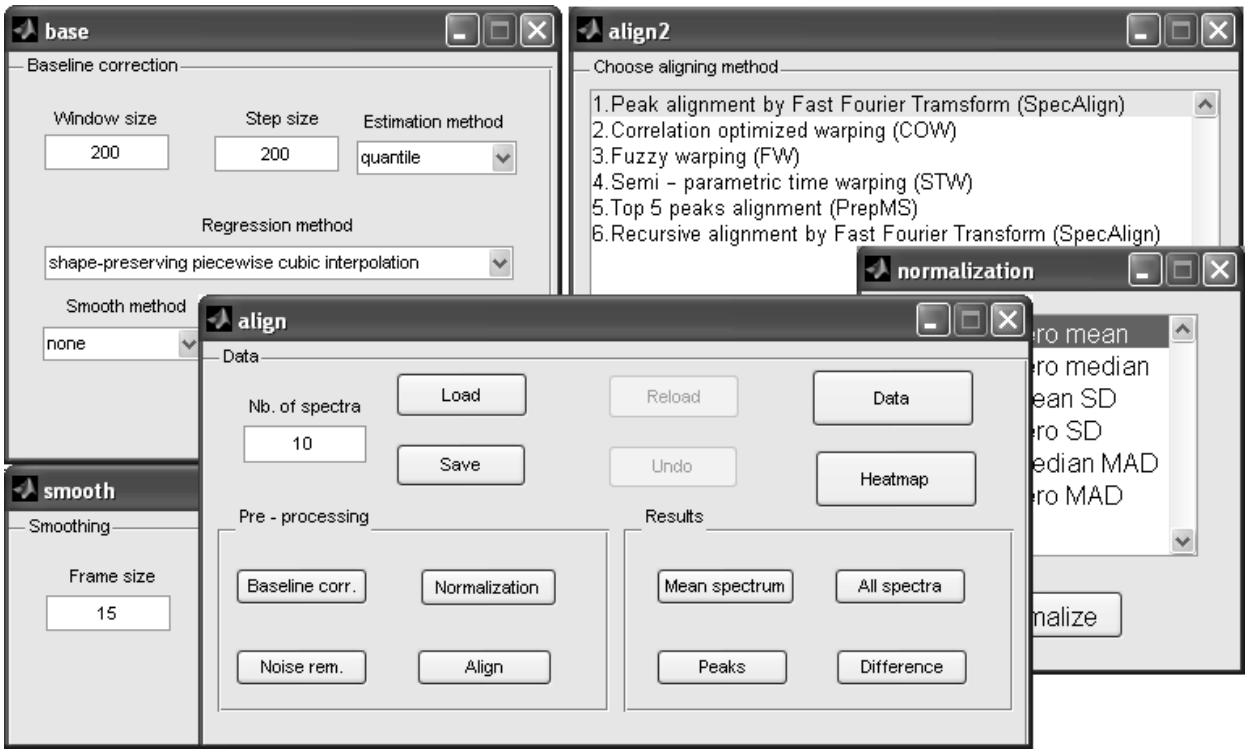

Fig.2. Align application snapshot. Main window in the center, baseline correction and alignment windows - top, smoothing window – bottom left and normalization window – bottom right.

4. Conclusions

Our application is a graphical tool, written in a commonly used Matlab environment, which provides a complete routine for mass spectra preprocessing. Align is user-friendly and advanced enough for the majority of scientists. Because no standard for mass spectra preprocessing has been established, large number of algorithms were implemented, which allows to choose the most optimal methods for different datasets. In every method, default parameters were used, which enables mass spectra preprocessing for an inexperienced users. However, all settings can be modified to satisfy requirements of the researcher, so it could be a very useful tool to prepare mass spectra for a further analysis.

This work was partially supported by BK 209/RAU1/2008, t.3

Bibliography

[1] Aebersold R, Mann M: *Mass spectrometry – based proteomics*, Nature 2003, 422, 198-207

[2] Cairns DA, Thompson D, Perkins DN, Stanley AJ, Selby PJ, Banks RE: *Proteomic profiling using mass spectrometry - does normalising by total ion current potentially mask some biological differences?*, Proteomics 2008, 8, 21-27

[3] Coombes KR, Fritsche HA Jr., Clarke C, Chen J, Baggerly KA, Morris JS, Xiao L, Hung M, Kuerer HM: *Quality Control and Peak Finding for Proteomics Data Collected from Nipple Aspirate Fluid by Surface-Enhanced Laser Desorption and Ionization*, Clin. Chemi. 2003, 49,1615–23

[4] Coombes KR, Baggerly KA, Morris JS: *Pre-processing mass spectrometry data.* In Dubitzky W, Granzow M, Berrar, DP (Eds.) Fundamentals of data mining in genomics and proteomics, Springer 2007, New York, 79-99

[5] Cruz-Marcelo A, Guerra R, Vannucci M, Li Y, Lau CC, Man T-K: *Comparison of algorithms for pre-processing of SELDI-TOF mass spectrometry data*, Bioinformatics 2008, 24, 2129-2136

[6] Du P, Kibbe WA, Lin SM: *Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching*, Bioinformatics 2006, 22, 2059-65

[7] Eilers PHC: *Parametric time warping*, Anal. Chem. 2004, 76, 404-11

[8] Eilers PHC, Marx BD: *Flexible smoothing with B-splines and penalties*, Stat. Sci. 1996, 11, 89-121

[9] Karpievitch YV, Hill EG, Smolka AJ, Morris JS, Coombes KR, Baggerly KA, Almeida JS: *PrepMS: TOF MS data graphical preprocessing tool*, Bioinformatics 2007, 23, 64-65

[10] Li X, Gentleman R, Lu X, Shi Q, Iglehart JD, Harris L, Miron A: *Seldi-tof mass spectrometry protein data.* In Gentleman,R. et al. (eds) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Ch. 6, Springer 2005, New York, 91-109.

[11] Malyarenko DI, Cooke WE, Adam BL, Malik G, Chen H, Tracy ER, Trosset MW, Sasinowski M, Semmes OJ, Manos DM: *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*, Clinical Chemistry 2004, 51, 65-74.

[12] Meuleman W, Engwegen J, Gast M, Beijnen J, Reinders M, Wessels: *Comparison of normalisation methods for Surface-Enhanced Laser Desorption and Ionisation Time-Of-Flight Mass Spectrometry data*, BMC Bioinformatics 2008, 9

[13] Monchamp P, Andrade-Cetto L, Zhang JY, Henson R: *Signal Processing Methods for Mass Spectrometry*. In Alterovitz G, Ramoni MF (Eds.) Systems Bioinformatics: An Engineering Case-Based Approach, Artech House Publishers 2007

[14] Morris JS, Coombes KR, Koomen J, Baggerly KA, Kobayashi R: *Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum,* Bioinformatics 2005, 21, 1764–1775

[15] Nielsen NP, Carstensen JM, Smedsgaard J: *Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimized warping*, J. Chromatogr. A 1998, 805, 17-35

[16] Paweletz CP, Trock B, Pennanen M, Tsangaris T, Magnant C, Liotta LA, Petricoin III EF: *Proteomic patterns of nipple aspirate fluids obtained by SELDI–TOF: potential for new biomarkers to aid in the diagnosis of breast cancer*, Dis. Markers 2001, 17, 301–307.

[17] Rifai N, Gillette MA, Carr SA: *Protein biomarker discovery and validation: the long and uncertain path to clinical utility*, Nature Biotechnology 2006, 24, 971-83

[18] Savitzky A, Golay MJE: *Smoothing and differentiation of data by simplified least squares procedures*, Anal. Chem. 1964, 36, 1627-39

[19] Sinkhorn R: *A relationship between arbitrary positive matrices and doubly stochastic matrices*, The Annals of Mathematical Statistics 1964, 35, 876-79

[20] Walczak B, Wu W: *Fuzzy warping of chromatograms*, Chemometrics and Intelligent Laboratory Systems 2005, 77, 173-80

[21] Wong JWH, Durante C, Cartwright HM: *Application of Fast Fourier Transform Cross-Correlation for the Alignment of Large Chromatographic and Spectral Datasets,* Anal. Chem. 2005, 77, 5655-61

[22] Wong J, Cagney G, Cartwright HM: *Specalign-processing and alignment of mass spectra datasets*, Bioinformatics 2005, 21, 2088–90.

[23] Yang C, He Z, Yu W: *Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis*, BMC Bioinformatics 2009, 10

Author:

MSc. Michal Marczyk
Silesian University of Technology
Akademicka 16
44-100 Gliwice
tel. 604 176 710
email: *Michal.Marczyk@polsl.pl*