Talanta 84 (2011) 1180–1187

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

A discriminant based charge deconvolution analysis pipeline for protein profiling of whole cell extracts using liquid chromatography–electrospray ionization-quadrupole time-of-flight mass spectrometry

Weiying Lu[a], John H. Callahan[b], Frederick S. Fry[b], Denis Andrzejewski[b], Steven M. Musser[b], Peter de B. Harrington[a,*]

[a] *Center for Intelligent Chemical Instrumentation, Clippinger Laboratories, Department of Chemistry and Biochemistry, OHIO University, Athens, OH 45701-2979, USA*
[b] *Instrumentation and Biophysics Branch, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, 5100 Paint Branch Parkway, College Park, MD 20740, USA*

ARTICLE INFO

Article history:
Received 29 October 2010
Received in revised form 7 March 2011
Accepted 10 March 2011
Available online 21 March 2011

Keywords:
Liquid chromatography–electrospray ionization mass spectrometry
Fuzzy rule-building expert system
Molecular weight determination charge state deconvolution
Discriminant based charge deconvolution
Salmonella enterica serovar typhimurium

ABSTRACT

A discriminant based charge deconvolution analysis pipeline is proposed. The molecular weight determination (MoWeD) charge deconvolution method was applied directly to the discrimination rules obtained by the fuzzy rule-building expert system (FuRES) pattern classifier. This approach was demonstrated with synthetic electrospray ionization-mass spectra. Identification of the tentative protein biomarkers by bacterial cell extracts of *Salmonella enterica* serovar typhimurium strains A1 and A19 by liquid chromatography–electrospray ionization-mass spectrometry (LC–ESI-MS) was also demonstrated. The data analysis time was reduced by applying this approach. In addition, this method was less affected by noise and baseline drift.

1. Introduction

Electrospray ionization-mass spectrometry (ESI-MS) methods such as liquid chromatography–electrospray ionization-mass spectrometry (LC–ESI-MS) and capillary electrophoresis–electrospray ionization-mass spectrometry (CE–ESI-MS) [1] have been applied to the analysis of intact proteins in recent years. CE–ESI-MS has been applied to the diagnosis of cancer [2], to detect glycoproteins [3] and ribosomal proteins of *Escherichia coli* [4], etc. Protein expression profiling analyses on bacteria whole cell extracts without proteolytic digestion by liquid chromatography–electrospray ionization-quadrupole time-of-flight mass spectrometry (LC–ESI-QTOF MS) have been reported [5–7]. A new analysis pipeline to process the LC–ESI-MS data is proposed for whole cell extracts. LC–ESI-MS spectra of biological samples often have hundreds of component peaks, and some peaks may a have low signal-to-noise ratio (SNR). The proposed analysis pipeline applies data processing methods including wavelet denoising, baseline removal, peak binning, peak centroiding, and multivariate classification to extract proteomic information. The analysis can be automated and the results are easier to reproduce than manual spectra inspection [5,8–10].

For biomolecules such as proteins, ESI-MS often produces multiply charged spectra. This multiplicity of charge states gives an envelope of peaks in a spectrum for each component. The charge state deconvolution method, also known as deconvolution, is a method that determines the molecular mass of a biomolecule from multiply charged ESI-MS peaks [11]. Deconvolution transforms a multiply charged ESI-MS spectrum into a zero-charge or singly charged spectrum. The zero-charge spectrum is a spectrum that each component data point has an abscissa of its molecular mass. Similarly, in the singly charged spectrum, each component data point has an abscissa that is the mass-to-charge ratio of its singly charged ion. All multiply charged spectra were deconvolved into zero-charge spectra in this work. Many deconvolution methods have been proposed, such as thorough high resolution analysis of spectra by Horn (THRASH) [12], molecular weight determination (MoWeD) [13], maximum entropy deconvolution [14], multiplicative correlation algorithm (MCA) [15], Zscore [16], etc. Generally, different deconvolution algorithms are designed for low-resolution and high-resolution ESI-MS data, based on whether the isotopic peaks are resolved [16]. MoWeD, MCA, and maximum entropy deconvolution were suitable for low-resolution mass spectrometry such as quadrupole time-of-flight mass spectrometry (QTOF-MS),

[*] Corresponding author.
E-mail address: Peter.Harrington@OHIO.edu (P.d.B. Harrington).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.03.024

which has unresolved isotopic peaks, particularly for high-charge states and high molecular weight proteins. THRASH was designed for high-resolution mass spectrometry (resolved isotopic peaks) such as Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS). Zscore has different deconvolution routines for each resolution. The deconvolution method was originally designed to deconvolve ESI-MS spectra. To deconvolve a two-way LC–ESI-MS spectrum, the deconvolution should be performed on the binned ESI-MS scan in a given time interval.

The deconvolved spectra can be used as input data to multivariate pattern classifiers, and the discriminant rules with candidate biomarker information can be obtained. In this work, this general approach was performed by the MoWeD deconvolution algorithm and the fuzzy rule-building expert system (FuRES) [17] as the pattern classifier. This general analysis approach is named MoWeD–FuRES. The MoWeD deconvolution algorithm was chosen because this algorithm is relatively efficient, simple, and is suitable for the deconvolution of low-resolution mass spectra. FuRES combines the advantage of the fuzzy logic data analysis with the decision tree algorithm. FuRES has been successfully applied in matrix-assisted laser desorption/ionization-mass spectrometry (MALDI-MS) data analysis such as mouse age identification [18], premature or at-term deliveries classifications of amniotic fluids [19]. FuRES is also capable of two-way data analysis in applications such as classification of jet fuels [20] and ignitable liquids [21].

In general analysis approaches such as MoWeD–FuRES, applying deconvolution and the data processing methods on an individual mass spectrum is time-consuming, especially when there are many samples and the noise threshold in the deconvolution method is low. A novel hyphenated approach named FuRES–MoWeD is proposed in the analysis pipeline, which applies charge deconvolution algorithm to multivariate pattern recognition rules. Because FuRES was performed prior to the deconvolution, only peaks correlated to each class are deconvolved. As a result, the FuRES–MoWeD approach is more robust in terms of the classification ability than MoWeD–FuRES with respect to baseline drift and noise. In addition, this approach is efficient because the deconvolution is performed only once for a set LC–ESI-MS of spectra. FuRES–MoWeD is compared to the MoWeD–FuRES approach on a synthetic data set and a *Salmonella enterica* strain identification data set.

2. Theory

The FuRES–MoWeD analysis approach focuses on effectively obtaining molecular component differences between different classes of samples. Fig. 1 gives the flowcharts of FuRES–MoWeD and MoWeD–FuRES approaches for comparison. Both approaches combine the pattern recognition methods with the charge deconvolution. In MoWeD–FuRES approach, charge deconvolution is applied to every ESI-MS scan. When the chromatographic separation time is long, the deconvolution will be computationally demanding. As a result, the deconvolution should be performed after binning the spectra in a given retention time interval.

In the proposed FuRES–MoWeD approach, FuRES classification is applied to the LC–ESI-MS spectra prior to charge deconvolution. The resultant FuRES discriminant comprises a two-way image of retention time and mass-to-charge ratio. The spectrum of each retention time in the discriminant is then deconvolved from the FuRES discriminant image. Because the discriminant image contains fewer data points than a complete set of sample data, FuRES–MoWeD approach shortens the analysis time compared to MoWeD–FuRES approach. In addition, because the discriminants only retain relevant information for the classification, there is less noise in the discriminants than the sample spectra. Therefore, the analysis result is less affected by noise in the spectra than MoWeD–FuRES approach.

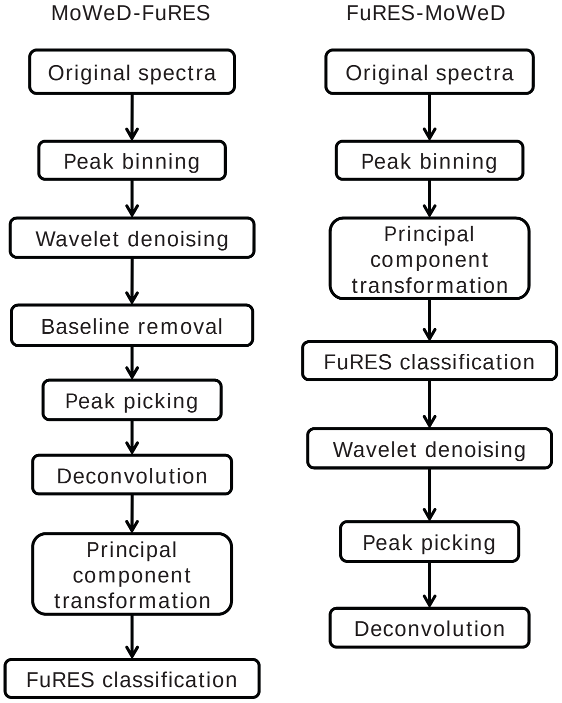

Fig. 1. Diagram of key steps in a general LC–ESI-MS analysis pipeline (MoWeD–FuRES) and the analysis pipeline proposed in this work (FuRES–MoWeD).

The MoWeD algorithm was proposed by Pearcy and Lee [13]. First, the peaks in the spectrum are detected by a peak-finding routine. Starting with the highest intensity peaks, this algorithm calculates and assigns charge states to each peak that is above a given noise threshold. After charge assignment of all peaks, the ESI-MS spectrum is deconvolved into the zero-charge domain. The charge assignment process contains the following steps:

1. Evaluate the range of candidate charges that is defined by the input mass-to-charge ratio range and the pre-defined maximum possible molecular mass.
2. For each candidate charge, calculate charge state patterns. Charge state patterns are a series of mass-to-charge ratio values from a contiguous charge state series.
3. A score is calculated for each candidate charge state by a scoring function. The scoring function is the number of peaks present in the charge state pattern minus the number of gaps in the charge state pattern.
4. The candidate charge with the maximum score is assigned to the peak.
5. The zero-charge spectrum is updated each time after a charge is assigned. For every point in an original peak, the molecular mass is calculated. The calculated molecular mass is the abscissa of the peak points and the peak intensity is the ordinate. According to the pre-defined molecular mass range and molecular mass increment, the intensities in the zero-charge spectrum are calculated by linear interpolation between adjacent peak points. The zero-charge spectrum is updated by adding the interpolated intensity.

Compared to a LC–ESI-MS spectrum, a FuRES discrimination rule is different because the discrimination rule contains both positive and negative peaks that respectively represent the relative importance for two classes. To extract the zero-charge spectrum from the FuRES discrimination rule, first the deconvolution was performed on the absolute value of the discrimination rule spectrum. Afterwards, the sign of each peak of the zero-charge spectrum from the deconvolved rule was determined by counting and comparing the number of positive and negative peaks in the charge state pattern calculated from the MoWeD algorithm (step 3). If the number of positive rule peaks in the pattern is larger than the negative peaks, the deconvolved rule peak is positive, and vice versa. The signs of the peaks in the zero-charge spectrum indicate the corresponding class. Positive peaks are selective for spectra that are partitioned to the right branch of the classification tree by the discrimination rule, and contrarily, the negative peaks are selective to the left branch of the classification tree.

3. Experimental

3.1. Synthetic data set

A synthetic data set with two classes was generated by methods adapted from the reference [22]. Class A comprised 30 synthetic ESI-MS spectra of horse heart myoglobin with baseline noise. The isotopic distribution of this protein was calculated by the polynomial algorithms [23] with a permutation threshold of 0.01. The charge distributions were calculated by the binomial distribution, assuming each basic amino acid in the protein has a probability of 0.5 to receive a proton. Each peak was a Gaussian peak with a full width at half maximum (FWHM) resolution of 10 000. The pure signal was normalized to a maximum intensity of 10. It is assumed that in LC–ESI-MS spectra of cell extracts, many chemical impurities irrelevant to identification will cause a bell shaped noise in the mass spectrometry domain. The noise model is further validated by Section 4.2. The chemical noise was simulated by a Gaussian function with amplitude of 30, a center of 0.4 and a standard deviation of 0.1 at the 0–1 abscissa. According to the literature [22], Poisson distributed shot noise was added. Each spectrum was sampled in a mass-to-charge ratio range of 550–2000 Th with an increment of 0.1 Th. Class B comprised 30 spectra containing baseline noise only. All spectra were normalized to unit length. The data set was stored in a 60 × 14 501 matrix, where the spectra were stored as rows.

3.2. Bacterium identification data set

Two strains, designated A1 and A19, of *Salmonella enterica* reference set A (SAR A) [24] were analyzed. Strains A1 and A19 are classified as part of the typhimurium serovar. All bacteria were grown 24 h on tryptic soy agar plates (Difco Laboratories, Sparks, MD). The cells were first vortexed to form a slurry of cells in 70% ethanol. Two hundred microliters of the slurry was collected and centrifuged to a pellet, and the 70% ethanol was removed. Proteins were extracted from bacterial cells with a 50:45:5 solution of acetonitrile (J.T. Baker, Phillipsburg, NJ), HPLC-grade water (J.T. Baker), and formic acid (Sigma–Aldrich, St. Louis, MO). The cells were mixed with 1 mL of the extraction solution and placed in a Barocycler extraction tube (Pressure Biosciences Inc., Boston, MA). In the Barocycler, the sample is exposed to cycles of high (35 kpsi) and low (0 kpsi) pressure. Each pressure is maintained for 25 s. A series of 10 cycles was performed to extract the proteins. The cellular debris was then centrifuged to a pellet, and the clear solution extract was removed.

Separations of protein extracts were performed on an Agilent 1100 HPLC system (Agilent Technologies, Palo Alto, CA) installed with two 150 mm × 2.1 mm Prosphere P-HR (Alltech Associates, Deerfield, IL) columns. The columns were sequentially connected to improve the chromatographic resolution. Mobile phase A and B was respectively 5% acetic acid in water, and 5% acetic acid in acetonitrile. After sample injection, the solvent composition was held constant at 10% B for 5 min, linearly increased to 50% B between 5 and 70 min, followed by a linear increase to 90% B at between 70 and 80 min. The solvent composition was then linearly changed back to 10% B by 110 min. The flow was split after the column with approximately 25% of the flow going to the mass spectrometer while the remaining eluent was diverted to an HP 1100 fraction collector.

Mass spectra were acquired on a Waters Q-Tof Premier mass spectrometer (Waters, Milford, MA) over a mass-to-charge ratio range of 550–2000 Th using electrospray ionization in the positive ion mode. The scan time was 2.0 s with a 0.1 s interscan delay. The spectra were collected in continuum mode. Five replicates were obtained for each bacterial strain.

The LC–ESI-MS spectra of *Salmonella enterica* were converted into ASCII text files using the Databridge program with MassLynx version 4.0 (Waters, Milford, MA). The text files were then imported into MATLAB. The original spectra were binned by a mass-to-charge ratio increment of 0.1 Th and a retention time increment of 0.1 min. The mass-to-charge ratio cutoff range was 550–2000 Th and the run time cutoff was 85 min. The MS scans were stored as rows in the matrix. Each binned spectrum was stored as an 851 × 14 501 matrix.

3.3. Data processing

All calculations were performed on a personal computer equipped with a Core i7 940 CPU and 12 GB memory running Microsoft Windows XP x64 SP2 operating system. All programs were in-house scripts written in MATLAB version 7.11 (The MathWorks Inc., Natick, MA). The MATLAB code used in this study are available upon request from the corresponding author.

In MoWeD–FuRES approach, each ESI-MS scan was denoised by a wavelet denoising method, in which the nonlinear discrete wavelet transform is applied. In FuRES–MoWeD approach, the same denoising method was applied to the FuRES discrimination rule. Wavelet denoising was performed by using the ThreshWave function in the WaveLab toolbox for MATLAB version 8.5 [25]. The wavelet filter used was the Symmlet level 4. The threshold was determined by the visually calibrated adaptive smoothing (VisuShrink) method [26]. Soft thresholding technique was applied on the wavelet coefficients. After applying wavelet denoising, the signal and noise components were separated. The wavelet based noise estimation spectra were then used to calculate the SNR for each peak.

The component peaks were identified by finding local maxima in the spectrum with SNRs larger than 3. The range of the component peak is defined by a starting point and an ending point. The starting point and ending point are two local minima of the spectrum, which are nearest to the peak maxima on both sides. When some peaks with low intensities appeared near a large peak and none of these neighboring peaks were more than three times more intense than the large peak, these small peaks were considered as the post-translational modification of this protein. Therefore, the same charge states were assigned to these small peaks.

In MoWeD–FuRES approach, the spectra were baseline corrected by using an iterative 10th order polynomial fitting after wavelet denoising. The iteration stopped when the number of the fitted points was less than 10% of the number of points in the spectrum or the median of the absolute value of the residuals converged.

The MoWeD charge deconvolution method was applied on the processed spectra and the FuRES discrimination rule for MoWeD–FuRES and FuRES–MoWeD approaches, respectively. The maximum possible molecular mass was 50 kDa. After deconvolution, the mass spectra were transformed to 550–50 000 Da with a molecular mass increment of 1 Da. For the bacterium identification data set, each deconvolved two-way spectrum was stored as an 851 × 49 451 matrix.

Each two-way object in the bacterium identification data set was unfolded to a vector that respectively had 42 082 801 and 12 340 351 points for the MoWeD–FuRES and FuRES–MoWeD approach. All vectors were normalized to unit length. All unfolded vectors were stored as rows in a 10 × 42 082 801 sparse matrix and a 10 × 12 340 351 sparse matrix for MoWeD–FuRES and FuRES–MoWeD approach, respectively. The principal component transformation (PCT) was applied as a lossless compression method before FuRES modeling. After PCT compression, the number of variables equals to the number of objects in the training set. Therefore, the computation time and memory requirement of the FuRES classification were greatly reduced.

4. Results and discussion

4.1. Synthetic data set

The examples of generated synthetic ESI-MS spectra were demonstrated in Fig. 2. The peak of horse heart myoglobin could not be observed in class A because of relatively low concentrations compared with the baseline noise. Fig. 3 demonstrates the input vectors for MoWeD deconvolution and the FuRES discrimination rules from the synthetic data set by two processing methods. Because the pure signal is three times weaker than the baseline, the MoWeD–FuRES could not identify the protein because each protein feature was not extracted from noise when deconvolution was applied on the individual unprocessed spectra. The protein signals were identified as high frequency noise and deconvolved into a high molecular mass domain. However, the FuRES–MoWeD could generate the discrimination rules correctly because the protein features were correctly extracted from the whole set of spectra by FuRES discrimination rules, which made it possible to distinguish between the protein peaks and noise.

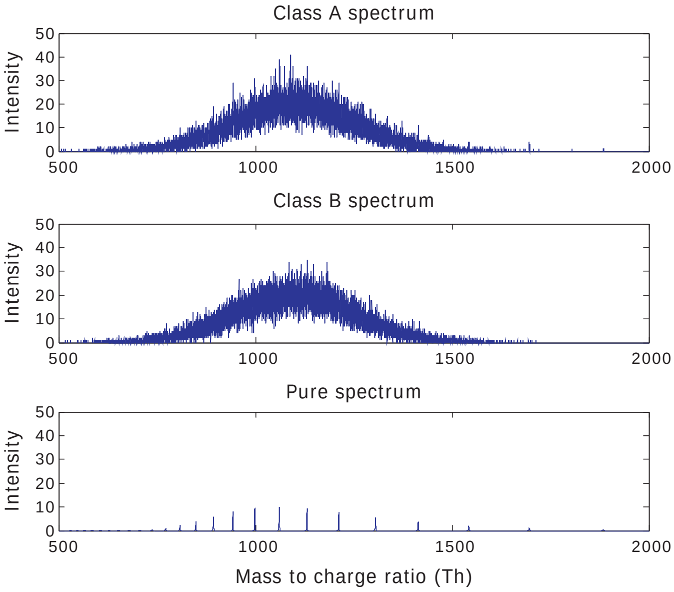

Fig. 2. Examples of synthetic horse heart myoglobin spectra (Class A and Class B) and the pure spectrum.

4.2. Bacterium identification data set

The example LC–ESI-MS spectra of two strains of *Salmonella enterica* are given in Fig. 4. The total ion chromatogram and total mass profile are also shown. The two-way spectra are transformed and plotted in logarithmic scale. From this figure, it is concluded that the two-way spectra of the two bacteria strains are similar. In addition, the noise components form a bell shape in the middle mass-to-charge ratio range and the retention time from

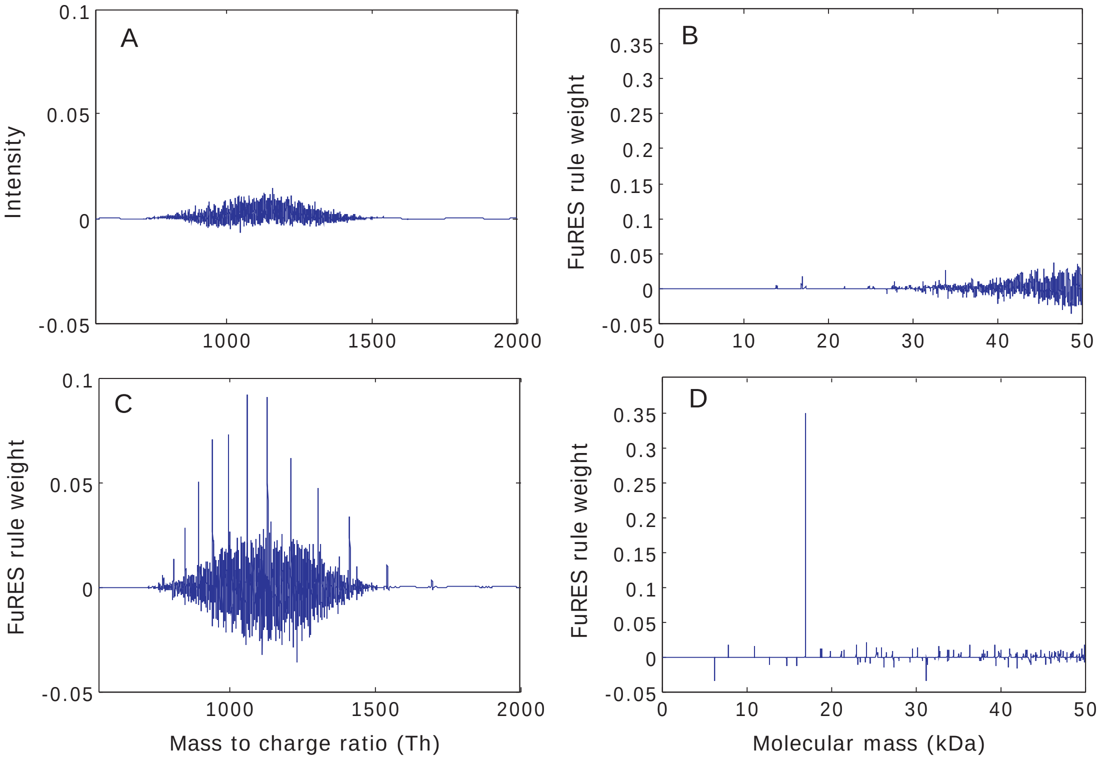

Fig. 3. The mass spectra and the discrimination rules of the synthetic data set processed by MoWeD–FuRES (top panels) and FuRES–MoWeD (bottom panels) approach. (A) An example of the denoised and baseline corrected spectrum as the input data for deconvolution, (B) the final MoWeD–FuRES discriminant rule, (C) the denoised and baseline corrected FuRES discriminant rule as the input data for deconvolution, (D) the final FuRES–MoWeD discriminant rule.

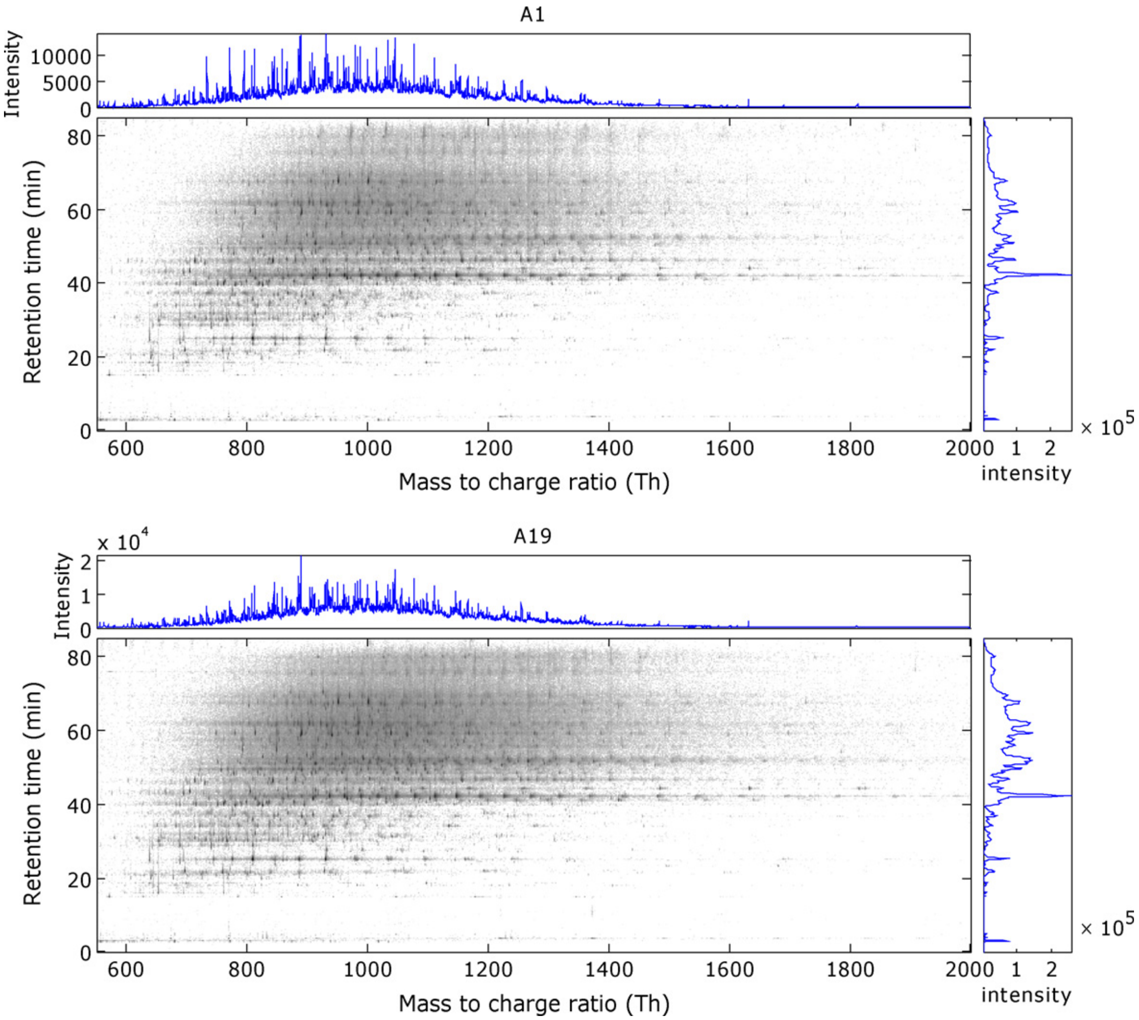

Fig. 4. Two-way LC–ESI-MS data objects of *Salmonella enterica* strains A1 and A19. The intensity in the two-way image is plotted in logarithmic scale to compare the amount of noise. The total ion chromatogram and total mass spectra are shown along with the two-way image.

20 to 80 min. This phenomenon is also observed from the total ion chromatogram and the total mass profile. The characteristics of noise in the experimental data were consistent with the noise simulated in the previous data set. The deconvolution results by MoWeD for a single spectrum of *Salmonella enterica* strains A1 and A19 are given in Fig. 5. The profiles generated by MoWeD were consistent with the spectra calculated by a commercial software package, ProTrawler 6 (BioAnalyte, Cambridge, MA). The comparison indicates the MoWeD deconvolution is a suitable method in this analysis pipeline for the bacterium identification data set. Because the deconvolution method used in ProTrawler is not publicly available, further comparisons were not performed.

When dealing with a two-way LC–ESI-MS data set, the data processing speed will be a more important factor compared to an ESI-MS data set, because the size of the data set is usually hundreds to thousands times larger. The effectiveness of the proposed FuRES–MoWeD approach is apparent in Table 1 by comparing run times and the total number of deconvolution routine evaluations. The actual computation time will vary by many factors such as the software and hardware configuration of computer system, the choice of programming language, etc. However, a general comparison of algorithmic efficiency between two approaches can be made

Table 1
Comparisons on run time and total number of deconvolution routine evaluations of bacterium identification data set.

	MoWeD–FuRES	FuRES–MoWeD
Total deconvolution routine evaluations	8510	851
Run time for deconvolution routine (min:s)	12:49	1:46
Total run time (min:s)	25:18	3:08

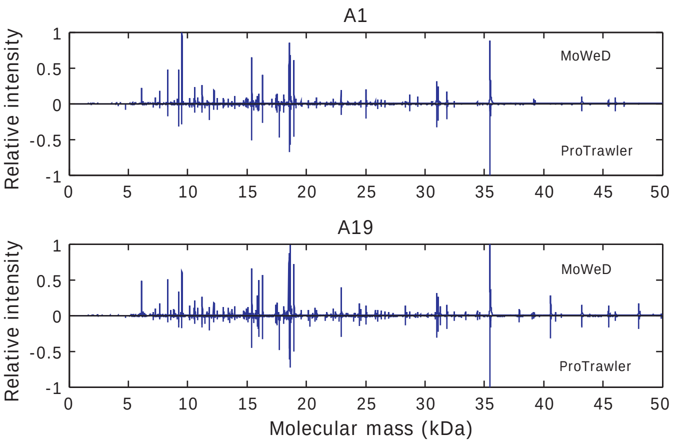

Fig. 5. Comparison of the deconvolved protein profiles by MoWeD and the ProTrawler on a representative spectrum of *Salmonella enterica* strains A1 (top panel) and A19 (bottom panel).

by program profiling. The run time was measured by the cputime function in MATLAB. Because the deconvolution calculation was performed only on FuRES discriminant, fewer deconvolution routine evaluations are needed. Moreover, the baseline correction routine is not performed. The FuRES–MoWeD approach required less time than the MoWeD–FuRES to perform.

The FuRES discrimination rules calculated by FuRES–MoWeD and MoWeD–FuRES approaches are given in Figs. 6 and 7. In Fig. 6, the discrimination rules were summed along the retention time dimension. The positive value indicates the corresponding component has characteristically high concentration in class A1 over A19, and vice versa. Similar summed spectra were obtained, which indicates the results of two comparing approaches are consistent in terms of detecting major proteomic features. The protein signals in both low and high molecular weight range can be observed sufficiently, which means the proposed method improves the detectability of high MW proteins when incorporated with FuRES. LC–ESI-MS provides the molecular weight and retention time of the proteins, which is insufficient to identify proteins due to various effects such as post-translational modification, mass errors and isotopic distribution. Five most abundant peaks and the corresponding tentative protein search results by SwissProt/TrEMBL database [27,28] are listed in Table 2. After searching the database with organism keyword "*Salmonella typhimurium*", a list that contains 15 744 protein entries were exported. The theoretical average molecular weights were then calculated by Compute pI/Mw, which is a part of ExPASy proteomics tools [29]. The search result was a protein entry that has the least mass difference between the observed mass and the calculated mass. Four out of five proteins were matched for both approaches.

The noise in LC–ESI-MS spectra possibly comes from the polymer contaminants from column degradation and other baseline noise components. These noise components can affect the deconvolution routine because noise peaks can be mistaken as signal peaks, which is a typical case that creates high mass artifacts in the deconvolved spectra. Fig. 7 demonstrates the two-way LC–ESI-MS discrimination rules. The intensities were plotted in logarithmic

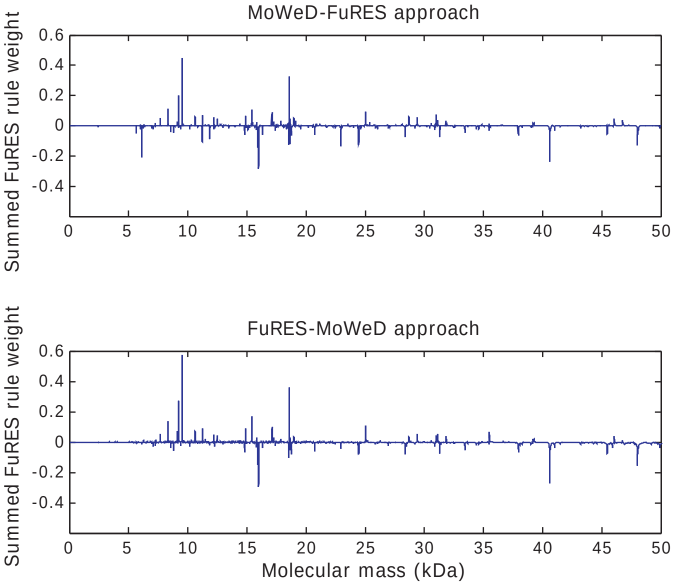

Fig. 6. Comparison of the summed discrimination rules calculated by MoWeD–FuRES (top panel) and FuRES–MoWeD (bottom panel) approaches. The discrimination rules are summed along the retention time dimension. The positive value indicates the corresponding component has characteristically high concentration in class A1 over A19, and vice versa.

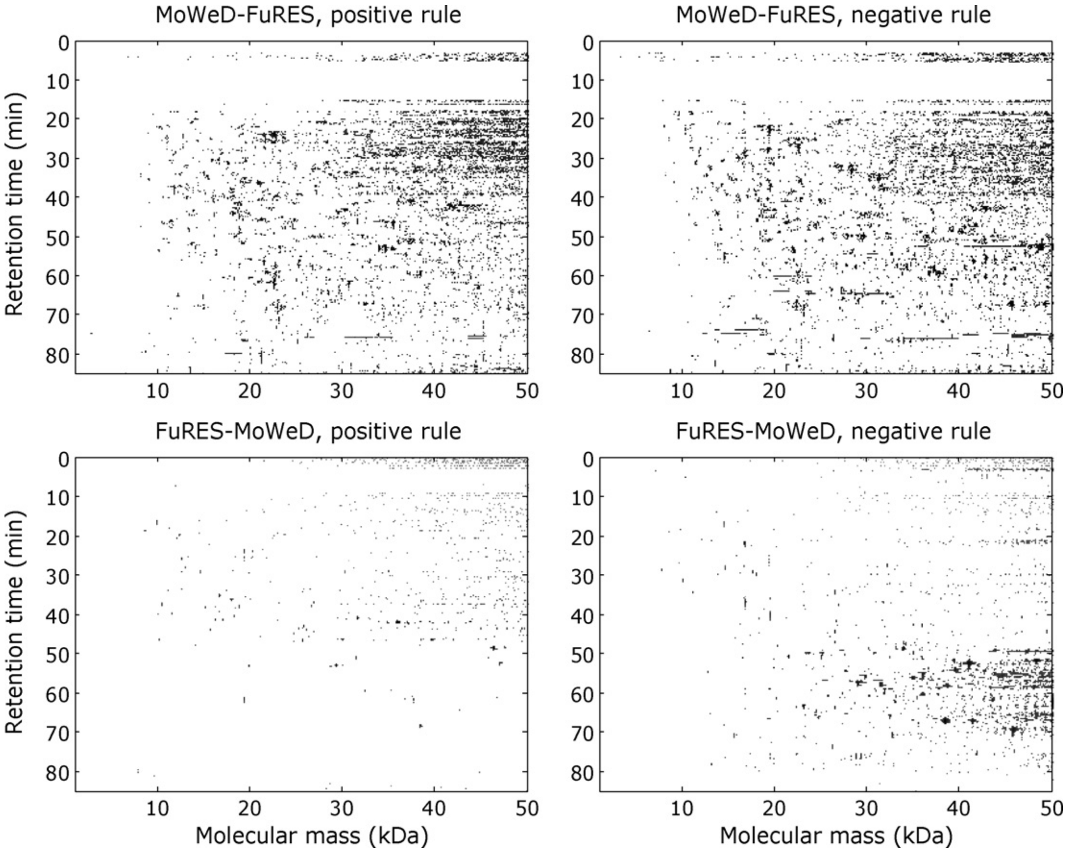

Fig. 7. The discrimination rules calculated by MoWeD–FuRES (top panels) and FuRES–MoWeD (bottom panels) approaches. The rule weights are plotted in logarithmic scale to compare the amount of noise.

Table 2
Five largest peaks observed in the summed discrimination rules and the tentative SwissProt/TrEMBL database search results for FuRES–MoWeD and MoWeD–FuRES approaches.

	Observed molecular mass (Da)	Calculated average molecular mass[a] (Da)	Relative intensity	Characteristic class[b]	Tentative accession ID
MoWeD–FuRES	9520	9520.96	1	A1	P0A1R6
	18,585	18586.08	0.7316	A1	Q7CQV9
	15,990	15988.81	0.6395	A19	P0A1J7
	40,595	40595.34	0.5315	A19	P19576
	6092	6094.94	0.4738	A19	D0ZSD2
FuRES–MoWeD	9520	9520.96	1	A1	P0A1R6
	18,585	18586.08	0.6270	A1	Q7CQV9
	15,990	15988.81	0.5171	A19	P0A1J7
	40,595	40595.34	0.4764	A19	P19576
	9239	9239.61	0.4704	A1	P0A1R8

[a] Results were obtained by Compute pI/Mw, which is a part of ExPASy proteomics tools.
[b] Characteristic class means the protein in the specified strain has relatively high concentration against the other strain.

scale to compare the amount of noise. It can be observed that there is a greater amount of noise in the rules obtained by the MoWeD–FuRES approach in the high mass range than that from the FuRES–MoWeD approach. The noise in the rules image is consistent with the results from the synthetic data set and previous study [5]. These artifacts were retained by using MoWeD–FuRES. In the FuRES–MoWeD approach however, noise levels are reduced in the FuRES discrimination rules before deconvolution, because the noise did not correlate with the systematic differences between the two bacterial strains. As a result, there are fewer artifact peaks in the FuRES–MoWeD approach than MoWeD–FuRES.

Theoretically, the FuRES classification rule is generated to minimize the entropy of classification between different classes, so that relevant protein signals are extracted for deconvolution without the cost of sensitivity and specificity. When and only when a protein signal is not characteristic for differentiating between the two rule consequents, that signal will not be present in the rule and thus be omitted from deconvolution. In practice, low-lying signals were correctly identified in the simulated data set. Additionally, the locations of low-intensity peaks were consistent between the two approaches in Fig. 6. The reason is that the FuRES rule looks for correlation among features of the spectra. There is a signal averaging benefit among the multiple charge state peaks and among objects common to each rule consequent.

Another concern is about the treatment of multiple charge variants of a protein, because the multiple charge envelope of a protein is different between different runs. Because the FuRES classification rule only picks signals with systematic differences between samples, the random variation of the multiple charge distribution between different samples will be omitted. On the other hand, the FuRES rules do treat systematic multiple charge variants as separate signals. However, a systematic change of variations between classes could possibly be caused by conformational changes [30], when the conformation and envelopes are on opposing consequents of the rule. Although the molecular weight remains unchanged, the proposed approach retains the specific proteins as tentative biomarkers in this regard. Again, further structure elucidation work is required for confirmation.

The FuRES–MoWeD uses the unprocessed spectra as training data, while the MoWeD–FuRES uses the deconvolved spectra. Therefore, the FuRES models obtained from these two methods have different predictive powers. Because both approaches apply pattern recognition methods during processing, internal validation methods can be applied. Bootstrapped Latin partition validation [31] was applied to compare the FuRES classification ability on the deconvolved spectra and the unprocessed spectra. The number of bootstraps and the number of partitions were respectively 10 and 2. The prediction results are in Table 3. The prediction accuracy was 89 ± 5% and 100% for MoWeD–FuRES and FuRES, respectively. This lower prediction accuracy indicates the loss of information in wavelet denoising, baseline removal and peak picking. The difference of the prediction results is statistically significant by two-way analysis of variance (ANOVA) with interaction at a significance level of 0.05.

Table 3
The confusion matrix of average correctly predicted objects with 95% confidence intervals between the two approaches of FuRES models. Each class contains five data objects.

	MoWeD–FuRES		FuRES	
	A1	A19	A1	A19
A1	4.6 ± 0.4	0.4 ± 0.4	5	0
A19	0.7 ± 0.5	4.3 ± 0.5	0	5

5. Conclusions

The proposed FuRES–MoWeD approach could rapidly find the features in complex sets of ESI-MS data. This approach was demonstrated by a synthetic ESI-MS data set and a LC–ESI-MS data set for bacteria strain identification, with the comparison of the MoWeD–FuRES approach. The resultant discrimination rule indicates that biomarker candidates can be found when signal to noise ratios in the spectra are low by FuRES–MoWeD approach. Performing the charge deconvolution on the FuRES discriminant rules as opposed to each individual across replicates yielded models less affected by baseline noise. In addition, the models obtained by two comparative approaches were evaluated by using bootstrapped Latin partitions. The performances of two approaches were evaluated with statistical measurements of confidence intervals.

This proposed approach was not limited to FuRES models and MoWeD methods demonstrated in the current study. Future work will involve the applications of deconvolution on rules of other classifiers such as partial least squares discriminant analysis, and using other deconvolution methods such as Zscore and MCA. This analysis pipeline is also feasible in the applications of high-resolution ESI-MS and CE–ESI-MS. The pipeline can be advantageous because the increased data size produced by high-resolution ESI-MS usually demands rapid analysis methods. As a pre-screening method, the proposed pipeline can be also applied in differential protein expression by liquid chromatography–tandem mass spectra (LC–MS/MS) that provide helpful fragmentation information to identify protein biomarkers specifically. Following the basic concept of the FuRES–MoWeD approach, the potential application can be extended further into other areas of proteomics. The processing method applied to the discriminant is not limited to charge deconvolution for LC–MS data. When processing LC–MS or LC–MS/MS profile, a database search routine for peptide fragments can be applied to the discriminant instead. Database searches on discriminants could potentially be useful when dealing with protein digest samples, where small peptide peaks are present.

Acknowledgements

Yao Lu, Xiaobo Sun, Zhanfeng Xu, and Zhengfang Wang are thanked for their helpful comments and suggestions.

References

[1] R. Haselberg, G.J. de Jong, G.W. Somsen, J. Chromatogr. A 1159 (2007) 81–109.
[2] D. Theodorescu, E. Schiffer, H.W. Bauer, F. Douwes, F. Eichhorn, R. Polley, T. Schmidt, W. Schofer, P. Zurbig, D.M. Good, J.J. Coon, H. Mischak, Proteomics Clin. Appl. 2 (2008) 556–570.
[3] A. Puerta, J. Bergquist, Electrophoresis 30 (2009) 2355–2365.
[4] M. Moini, H. Huang, Electrophoresis 25 (1981–1987).
[5] T.L. Williams, P. Leopold, S. Musser, Anal. Chem. 74 (2002) 5807–5813.
[6] R.A. Everley, T.M. Mott, S.A. Wyatt, D.M. Toney, T.R. Croley, J. Am. Soc. Mass Spectrom. 19 (2008) 1621–1628.
[7] T.M. Mott, R.A. Everley, S.A. Wyatt, D.M. Toney, T.R. Croley, Int. J. Mass spectrom. 291 (2010) 24–32.
[8] J. Trygg, E. Holmes, T. Lundstedt, J. Proteome Res. 6 (2007) 469–479.
[9] S. Bijlsma, L. Bobeldijk, E.R. Verheij, R. Ramaker, S. Kochhar, I.A. Macdonald, B. van Ommen, A.K. Smilde, Anal. Chem. 78 (2006) 567–574.
[10] F.T. Michaud, A. Garnier, L. Lemieux, C. Duchesne, Proteomics 9 (2009) 512–520.
[11] M. Mann, C.K. Meng, J.B. Fenn, Anal. Chem. 61 (1989) 1702–1708.
[12] D.M. Horn, R.A. Zubarev, F.W. McLafferty, J. Am. Soc. Mass Spectrom. 11 (2000) 320–332.
[13] J.O. Pearcy, T.D. Lee, J. Am. Soc. Mass Spectrom. 12 (2001) 599–606.
[14] A.G. Ferrige, M.J. Seddon, B.N. Green, S.A. Jarvis, J. Skilling, Rapid Commun. Mass Spectrom. 6 (1992) 707–711.
[15] J.J. Hagen, C.A. Monnig, Anal. Chem. 66 (1994) 1877–1883.
[16] Z.Q. Zhang, A.G. Marshall, J. Am. Soc. Mass Spectrom. 9 (1998) 225–233.
[17] P.B. Harrington, J. Chemometr. 5 (1991) 467–486.
[18] P.B. Harrington, C. Laurent, D.F. Levinson, P. Levitt, S.P. Markey, Anal. Chim. Acta 599 (2007) 219–231.
[19] P.d.B. Harrington, N.E. Vieira, P. Chen, J. Espinoza, J.K. Nien, R. Romero, A.L. Yergey, Chemometr. Intell. Lab. Syst. 82 (2006) 283–293.
[20] P. Rearden, P.B. Harrington, J.J. Karnes, C.E. Bunker, Anal. Chem. 79 (2007) 1485–1491.
[21] Y. Lu, P.B. Harrington, Anal. Chem. 79 (2007) 6752–6759.
[22] O. Schulz-Trieglaff, N. Pfeifer, C. Gropl, O. Kohlbacher, K. Reinert, BMC Bioinformatics 9 (2008).
[23] D.B. Hibbert, Chemometr. Intell. Lab. Syst. 6 (1989) 203–212.
[24] P. Beltran, S.A. Plock, N.H. Smith, T.S. Whittam, D.C. Old, R.K. Selander, J. Gen. Microbiol. 137 (1991) 601–606.
[25] D. Donoho, M. Duncan, X. Huo, O. Levi-Tsabari, WAVELAB 850, http://www-stat.stanford.edu/~wavelab/ (accessed December 2009).
[26] D.L. Donoho, J.M. Johnstone, Biometrika 81 (1994) 425–455.
[27] The UniProt Consortium, Nucleic Acids Res. 38 (2009) D142–D148.
[28] E. Jain, A. Bairoch, S. Duvaud, I. Phan, N. Redaschi, B. Suzek, M. Martin, P. McGarvey, E. Gasteiger, BMC Bioinformatics 10 (2009) 136.
[29] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, A. Bairoch, in: J.M. Walker (Ed.), The Proteomics Protocols Handbook, Humana Press, Totowa, New Jersey, USA, 2005, pp. 571–607.
[30] L. Konermann, B.A. Collings, D.J. Douglas, Biochemistry 36 (1997) 5554–5559.
[31] P.B. Harrington, Trends Anal. Chem. 25 (2006) 1112–1124.