A Time Series Approach for Clustering Mass Spectrometry Data[☆]

Francesco Gullo[a], Giovanni Ponti[a], Andrea Tagarelli[a,*], Giuseppe Tradigo[b], Pierangelo Veltri[b]

[a]DEIS, University of Calabria
via P. Bucci 41C
87036 Rende, Italy
[b]Bioinformatics Laboratory
Experimental and Clinical Medicine Department
Magna Græcia University of Catanzaro
viale Europa 88100
Catanzaro, Italy

Abstract

Advanced statistical techniques and data mining methods have been recognized as a powerful support for mass spectrometry (MS) data analysis. Particularly, supporting biomarker discovery in protein profiles has drawn special attention from biologists as well as clinicians. Moreover, due to its unsupervised learning nature, data clustering methods and algorithms have attracted increased interest for exploring, identifying, and discriminating pathological cases from MS clinical samples. However, the huge amount of information contained in a single sample (i.e., high-dimensionality), still limits the reliability of automatic information discovery in the effective identification of biomarkers.

In this paper, we present a data mining framework for MS data, in which the mining phase is focused on the task of clustering. Under the natural assumption of modeling MS spectra as time series, we propose a new representation model of MS spectra which allows for significantly reducing the high-dimensionality of such data, while preserving the relevant features. Besides the reduction of high-dimensionality (which typically affects effectiveness and efficiency of computational methods), the proposed representation model of MS data also makes the critical task of preprocessing the raw spectra less relevant in the whole process of MS data analysis. We evaluated our MS data clustering framework to publicly available proteomic datasets, and experimental results have shown the effectiveness of the proposed approach that can be used to aid clinicians in studying and formulating diagnosis of pathological states.

Keywords: clinical data, proteomics, mass spectrometry, clustering, time series

Introduction

Mass spectrometry (MS) is a powerful technique aimed at identifying and quantifying macromolecules (proteins) from biological samples [1]. MS allows for detecting ions within the input samples having different mass from each other. The result is a *mass spectrum*, i.e., a (large) sequence of (*m/z*, *I*) pairs which represent the *mass-to-charge ratio* (*m/z*) versus the quantity of macromolecules detected (*I*). Any mass spectrum corresponds to plotting the intensity values associated to fragments of sample compounds, as shown in Fig. 1. Mass spectra are generated by employing techniques that usually differ in such aspects as, e.g., the tuning of various parameters concerning sample preparation and ionization, the spectrometer type, the process used for detecting ions/macromolecules. Two of the earliest techniques for generating mass spectra are *Matrix-Assisted Laser Desorption/Ionization - Time Of*

[☆] An abridged version of this paper appeared in *Proceedings of the 20th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pages 412–417, 2007.

[*] Corresponding Author. Phone: +39 0984 494751, fax: +39 0984 494713.

Email addresses: fgullo@deis.unical.it (Francesco Gullo), gponti@deis.unical.it (Giovanni Ponti), tagarelli@deis.unical.it (Andrea Tagarelli), gtradigo@unicz.it (Giuseppe Tradigo), veltri@unicz.it (Pierangelo Veltri)

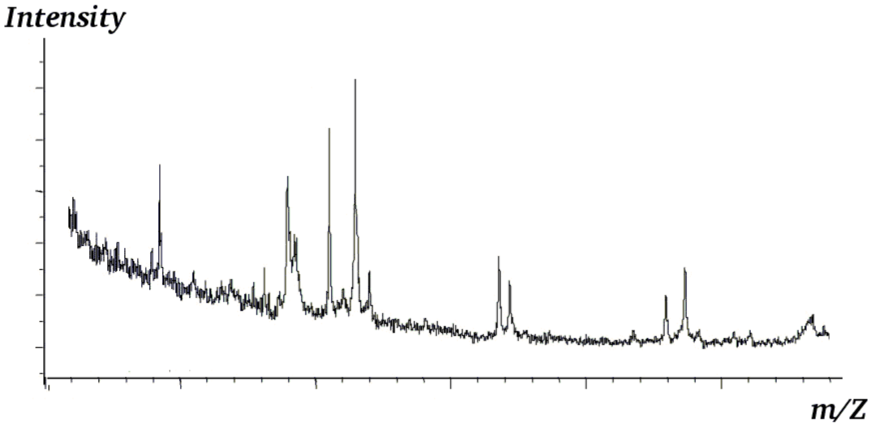


Figure 1: Graphical representation of a mass spectrum

Flight Mass Spectrometry (MALDI-TOF MS) [2] and *Surface-Enhanced Laser Desorption/Ionization - Time Of Flight Mass Spectrometry* (SELDI-TOF MS) [3]. Although MALDI-TOF and SELDI-TOF have been widely used with success for years [4], a number of novel techniques for MS generation have been developed recently, such as *Liquid Chromatography ElectroSpray Ionisation Mass Spectrometry* (LC-ESI MS) [5] or *Capillary Electrophoresis Mass Spectrometry* (CE MS) [6].

The identification of proteins contained in the original samples is performed by analyzing evolutions of the spectrum plot and selecting specific portions (peaks) of the spectrum used to query publicly available databases storing information about intensity expressions of known macromolecules (proteins). Unfortunately, the *high-dimensionality* of the data produced (i.e., the very large number of (*m/z*, *I*) pairs that typically compose mass spectra), as well as the strong influence of noisy factors, makes the process of protein identification very hard to perform. For this purpose, MS has been recently coupled with advanced automatic data analysis and mining techniques which aim to help clinicians for early detecting disease-related biological states. In this respect, a commonly used task consists in discriminating spectra based on the individual biological states (e.g., healthy or diseased) [7, 8]. This is performed by the data mining task of *supervised classification* or simply *classification*, whose main goal is to learn a mathematical model from a training set of positive/negative (i.e., healthy/diseased) examples whose classification is known; such a model is eventually exploited for recognizing the class (positive or negative, i.e., healthy or diseased) of unknown instances (spectra). Although classification of MS data has attracted the attention of many researchers [9, 10, 10–14], this task cannot be successfully applied when a training set of pre-classified instances is scarcely (or not at all) available. Also, building a training set requires careful human intervention, which limits the automation of the whole task. By contrast, the related data mining task of MS *clustering* consists in organizing a collection of spectra (whose classification is unknown) into meaningful groups (i.e., *clusters*), based on interesting relationships discovered from data (e.g., separating healthy individuals from diseased ones). Clustering finds natural application to many real MS scenarios, since the various pathologic states from clinical studies need to be identified and discriminated in an unsupervised way, i.e., without any training examples.

The aforementioned issues on MS analysis concerning high dimensionality and noise still affect such MS data analysis techniques as classification and clustering. In particular, most existing MS clustering approaches manage mass spectra in the classic count-vector-based form, which consists in treating the whole list of (*m/z*, *I*) pairs. Unfortunately, this leads to MS clustering methods that suffer from both high dimensionality, as a large number of (*m/z*, *I*) pairs) must be taken into account, and noise, which is inherently present in all mass spectra. Particularly, the problem of noise reduction in MS data is typically addressed in the preprocessing phase; nevertheless, MS preprocessing still represents a critical point in MS data analysis, due to the implicit difficulty in choosing the correct preprocessing steps to be performed, as well as their order of execution [15].

In this paper, we propose a framework for clustering MS data which has the main advantage of simplifying the MS preprocessing phase and alleviating the high dimensionality issue.

Figure 2 depicts the conceptual architecture of the proposed framework by showing its main constituent modules:

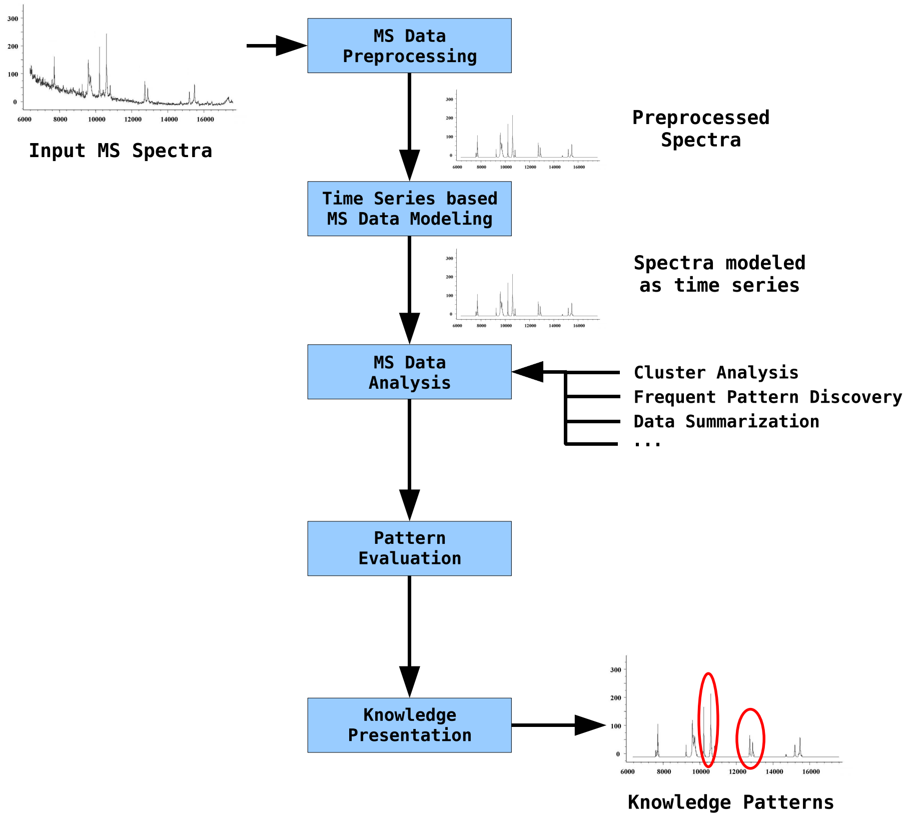


Figure 2: Conceptual architecture of the framework

(*i*) MS Data Preprocessing — it performs one or more preprocessing steps on the raw spectra in order to prepare them to the clustering task; (*ii*) Time Series based MS Data Modeling — this module transforms the preprocessed MS data into time series, preferably using a model conceived to maintain the significant trends (peak profiles) while reducing the data dimensions; (*iii*) MS Data Analysis — it performs one of the clustering tasks on the preprocessed spectra; (*iv*) Pattern Evaluation — is in charge of assessing the quality of clustering results, by showing discriminatory patterns found in the clusters; (*v*) Knowledge Presentation — highlights the discriminatory patterns found in data anche presents them in a readable form.

The key idea underlying the proposed framework is to cluster MS data by modeling the temporal information implicitly contained in mass spectra as a summarized *time series* based representation [16]. Representing mass spectra according to our time-series based model has the following main advantages with respect to the traditional count-vector based approach. First, the reduction of high dimensionality of MS data is accomplished by partitioning time series into variable-length segments , and then summarizing each segment by a numerical value. This also allows for drastically reducing the number of noisy dimensions (i.e., irrelevant (m/z, I) pairs) while preserving relevant ones (i.e., trends in the series profile). Second, the use of time series similarity detection techniques allows for simplifying the whole process of MS data analysis, thus that the raw spectra preprocessing phase becomes less relevant in providing reliable results. Experimental evaluations have been conducted on a number of large, publicly available MS datasets, to assess effectiveness of the proposed approach in clustering MS data, confirming that the proposed framework can be used supoprting biologists and clinicians in biomarker identification.

Related Work

Mass spectra analysis involves many activities aimed at preprocessing, identifying, classifying and clustering information obtained from different MS experiments. *Preprocessing* MS data has been recognized as a mandatory phase which aims to prepare spectra for further analysis [17]. The need for preprocessing mass spectra arises since such data are naturally corrupted by various noisy factors. Several research studies have been proposed on the development of preprocessing steps for MS data [8, 18], and in some cases they focused on specific steps, such as baseline subtraction [19–21], peak identification [22, 23], peak alignment [20, 24, 25]. In this respect, [26] studies the problem of prevent correlated variables by proposing a preprocessing algorithm based on the Bayes information criterion, which is exploited for selecting an optimal number of clusters. Biological markers are extracted from clinical SELDI-TOF mass spectra, which are derived from plasma by patients affected by Kawasaki disease and bone-marrow cells by patients affected by acute myeloid or acute lymphoblastic leukemia. [27] presents a novel geometry-based alignment algorithm working on LC-MS/MS data, i.e., MS data whose masses information is associated to chromatographic time. The work also uses the online available MSPTool [28] to perform well-known MS data preprocessing steps. In [29], peaks in mass spectra are identified according to a method that exploits data obtained by different MS instruments having different quality. That method does not involves any unsupervised approach to the identification of interesting peaks. In [30], a novel preprocessing model using quadratic variance is described for SELDI-TOF MS data. It has been demonstrated that such a model correctly characterizes the noise fluctuation observed in data by an exponential family of functions, thus allowing to have a better sensitivity in peak detection. In [31], a preprocessing software package called Wave-spec is described. It implements a number of methods for preprocessing MS data aimed at eliminating noise, aligning peaks among different spectra, detecting and quantifying peaks. Finally, it has been recently showed how wrongly choosing either the set of preprocessing algorithms/steps or their execution order might bias the biological interpretation of the study [15].

Several general purpose platforms for managing MS data have been developed [32–35]. In [32], a system based on the integration of different platforms for standalone applications is presented to compare healthy and diseased patients by using a statistical based approach based on the Wilcoxon-Mann-Whitney (WNM) algorithm. The experiments involve large MALDI-TOF datasets and make use of MS preprocessing steps, such as calibration and Fourier transformation. The classification of patients is eventually performed by using peptides contained in the original spectra. [33] presents the MAss SPECTRometry Analysis System (MASPECTRAS) as an integrated platform that allows to manage and analyze LC-MS/MS data. MASPECTRAS is based on the Proteome Experimental Data Repository (PEDRo). Its main functionalities include: gathering and parsing of the results from different search engines (such as SEQUEST, Mascot, etc.), peptide validation, clustering of proteins based on Markov clustering and multiple alignments, and quantification (using the Automated Statistical Analysis of Protein Abundance Ratios algorithm). Note that MASPECTRAS does not focus on clinical information extraction from MS data and does not perform any clustering task. The MsAnalyzer system [34] allows a graphical composition of workflows that describe the entire process of MS data analysis, including preprocessing. The graphical composition of workflows involves an ontology-based algorithm. [35] describes an integrated platform for preprocessing and mining MS data. The tools provided by that platform are: (i) preprocessing and data preparation, including peak identification, (ii) database querying for comparing the detected peaks to ones stored in publicly available databases (i.e., expressions of known proteins), (iii) MS clustering and classification.

Data mining techniques have been recognized as a valuable support to discovering significant patterns from MS data. The main focus in the MS domain has been on the task of classification, that is learning how to assign each instance with a category from a set of predefined categories (classes). The problem of MS classification has been addressed by using different machine learning techniques, including decision trees [9, 10], discriminant analysis [11], support vector machines [10, 12], and genetic algorithms [13, 14]. A comparison of different classification methods has been presented in [7, 8]. Like classification, clustering has been attracting a growing interest in various MS applications. A common way to address this problem is to apply classic clustering schemes and equip them with the Euclidean distance. For instance, in [36] an average-linkage agglomerative hierarchical clustering equipped with Euclidean distance is used in order to identify groups of proteins that show similar patterns. In [37], a two step hierarchical clustering of peak signals is applied to consolidate peak lists in multiple replicated experiments. That method is applied to peak lists of MS metabolic profiling data acquired from Leishmania mexicana, in order to separate the wild-type and two mutant parasite lines based on their metabolic profile. Principal component analysis

(PCA) technique is employed in [38] for comparing spectra focusing on peak heights. [39] uses a grid-based clustering algorithm to discover the functional molecules at the basis of the pharmacological compounds. In [40], mass spectra analysis is carried out by applying a discrete Fourier transformation method. A thresholding approach is adopted to denoise and reduce the length of each spectrum. A Bayesian model-based clustering algorithm is then applied to discover two groups of samples. That paper shows potential application of clustering for clinical MS data, thus demonstrating practical impact of using clustering in a MS context. Finally, clustering techniques are also used in MS/MS (tandem) data analysis, which is a different technique for biological sample analysis using two phase sample ionization. Clustering and merging spectra data produced in tandem spectrometry is proposed in [41]. That work presents a novel approach to calculating the similarity of fragmentation mass spectra based on increasing precision of modern mass spectrometry. The proposed algorithm is also applied to several proteomic datasets, giving an in-depth analysis of the influence of their parameters on the results. In [42], a consensus-based approach is exploited, in order to combine outcomes deriving from competing features ranking procedures. A consensus list is obtained and used to extract features from two publicly available protein datasets.

It should be noted that most approaches to clustering MS data mainly differ from each other about the preprocessing steps and the clustering scheme used, whereas spectra representation models have been rarely investigated. By contrast, this work is based on modeling spectra exploiting the notion implicitly present in MS data of temporal evolution of *m/z* values. This allows for exploiting a time series based model for representing mass spectra, which has the great advantage of effectively addressing the crucial issues concerning high dimensionality and noise within mass spectra. Also, most of the above works on MS data classification/clustering assume that clinical studies have been conducted on the data collections being available as a-priori knowledge on the data domain. This typically drives the development of classification/clustering methodologies that may focus on some portions only of the original spectra, that is portions which likely contain potential discriminatory patterns (biomarkers). By contrast, our approach can be in principle applied to the entire spectra as well.

Finally, this work can be considered as a complete work collecting modules previously published by the same authors; in particular [47] and [28] focus on the description of main software modules (preprocessing and data analysis) used for the complete study presented in this paper, while in [16] the DSA model is defined and its impact on similarity detection and clustering/classification of time series is thoroughly investigated.

Methods

Mass spectrometry techniques aim to generate *raw mass spectra*, which are (large) sequences of (*m/z*, *I*) pairs. Any raw mass spectrum can be viewed as a composition of three signals: the actual biological information, noise, and a baseline signal [18]. The latter is also referred to as *base intensity level*, as it varies across the *m/z* axis, so that intensity values that are under such baseline correspond to ground noise. Figure 3 shows an example of raw mass spectrum, where the three aforementioned signals are highlighted.

To extract useful information from raw spectra, both noise and baseline signals should be filtered out. This process, known as *MS data preprocessing*, is mandatory in order to make the signal containing biological information amenable to further analysis phases. Preprocessing may be performed depending on the MS platforms involved, experimental conditions and clinical studies [8, 18]. Moreover, selecting the proper set of specific preprocessing operations (i.e., steps) to be performed on raw mass spectra, as well as their execution order, generally depends on the goals of the task(s) of data analysis. Most preprocessing steps focus on the notion of *peak* in a mass spectrum. A peak is an intensity value corresponding to a local maximum such that its detection by the mass spectrometer has been performed for any ion fragment considering a unitary mass-to-charge ratio. Classic MS preprocessing steps include: *calibration* (i.e., mapping the observed time of flight into the inferred mass-to-charge ratio), *filtering or denoising* (i.e., reducing random noise) *baseline subtraction* (i.e., filtering out the baseline signal), *normalization* (i.e., making peak intensities understandable over a uniform range), *peak detection* (i.e., locating specific proteins or peptides across the *m/z* axis), *peak quantification* (i.e., representing detected peaks by means of a concise information), *peak matching/alignment* (i.e., recognizing the peaks in different spectra that correspond to the same biological macromolecule).

After preprocessing, mass spectra are modeled as *time series*. A *time series* is a sequence of real numerical values upon which a total order based on timestamps is defined. Time series are generally used to represent the temporal evolution of objects, hence huge amounts of such data are naturally available from several sources from different

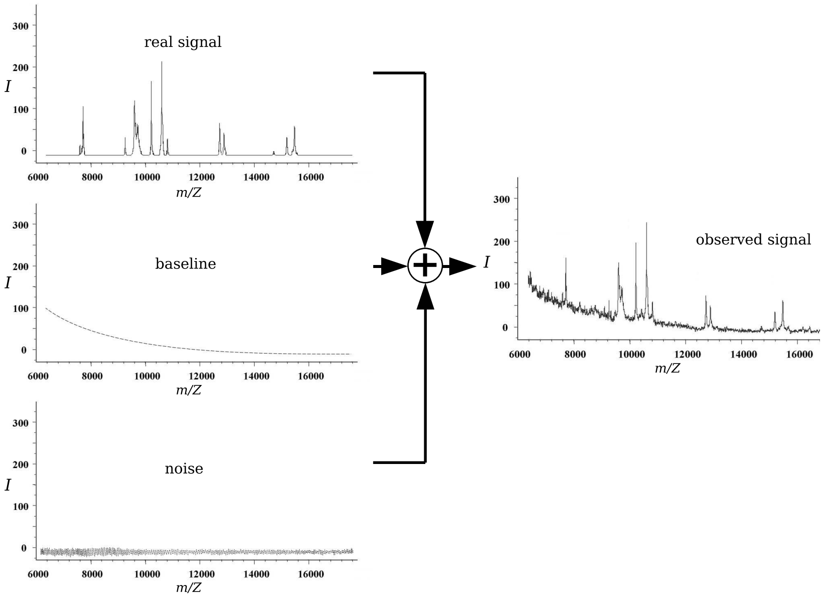


Figure 3: A raw mass spectrum

domains, such as speech recognition, biomedical measurement, financial and market data analysis, telecommunication and telemetry, sensor networking, motion tracking, and meteorology [43].

A dimensionality reduction step is performed on mass spectra modeled as time series. This is carried out by applying the *Derivative time series Segment Approximation* (DSA) representation model [16], which provides a concise, yet feature-rich representation of time series data. At the end of this step, each original mass spectrum is modeled as a *DSA sequence*, i.e., a time series represented according to the DSA model.

Most research on time series data management and knowledge discovery has been devoted to the similarity search and detection problem, which arises in many tasks such as indexing and query processing, change detection, classification, and clustering [43]. A common and effective method for comparing time series data to each other consists in "warping" the time axis in order to achieve the best alignment between the data points within the series to be compared. Such an alignment is discovered by means of the Dynamic Time Warping (DTW) algorithm [44]. Unlike the Euclidean distance, which performs a point-to-point comparison, DTW allows elastic shifting of a sequence to provide a better match with another sequence, thus it can handle time series (i.e., mass spectra) with local time shifting and different lengths. Figure 4 shows an example of two time series aligned according to (a) the Euclidean distance and (b) the DTW. Comparing mass spectra represented as time series by means of DTW makes some preprocessing steps unnecessary, as they are automatically accomplished by DTW. As an example, by exploiting the capability of handling local shifting, DTW implicitly performs the crucial MS preprocessing step of peak alignment. In this way, the complexity of making critical choices about MS preprocessing (i.e., which steps in what order) is mitigated, thus reducing the hardness of the entire preprocessing task and supporting one of the claims of this work.

Mass spectra modeled as DSA sequences are eventually involved into a task of clustering, in order to discover homogeneous, previously unknown groups so that the spectra within each group share common characteristics and/or relationships with one another. Mass spectra (DSA sequences) are clustered by using a standard centroid-based partitional clustering approach [45], where the similarity is computed according to the DTW algorithm.

The proposed framework can be summarized as follows:

1. MS data preprocessing applied to raw mass spectra;
2. time series based MS data modeling, i.e., transforming the preprocessed MS data into time series, in order to exploit the DSA time series representation model;

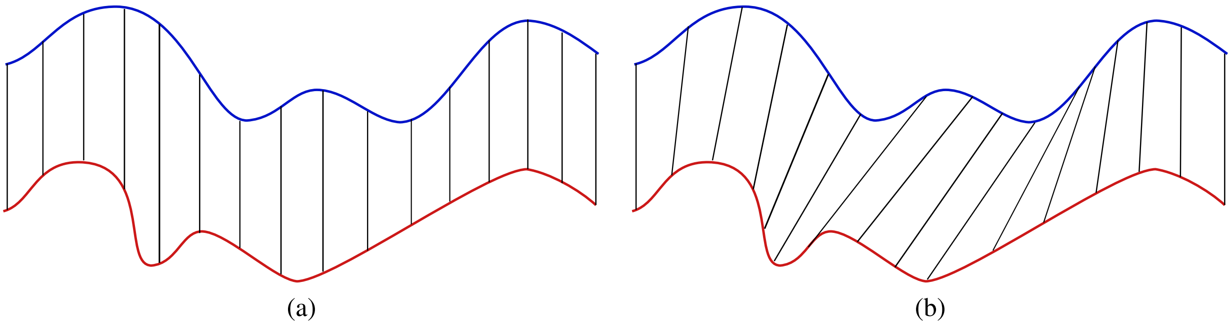


Figure 4: Time series alignment performed by (a) the Euclidean distance and (b) the Dynamic Time Warping

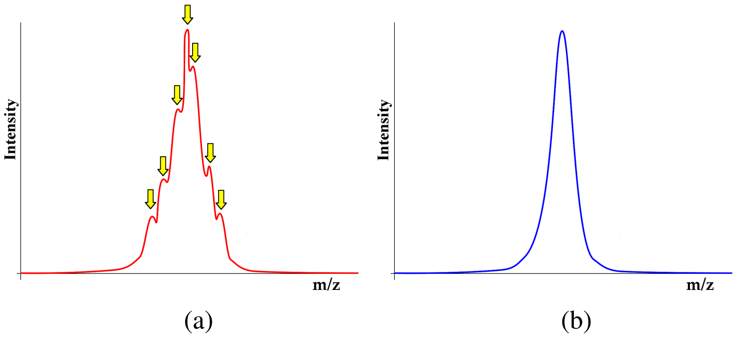


Figure 5: (a) Example M-peaks and (b) the corresponding ideal peak

3. MS data clustering and clustering evaluation to assess the quality of clustering results, by showing discriminatory patterns found in the clusters

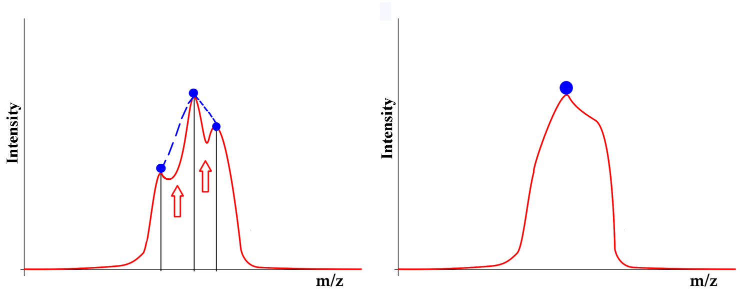
We explain the details of each part of the proposed framework next.

Preprocessing

MS data preprocessing is performed by applying standard operations [18]:

- Peak intensities are normalized using min-max normalization (i.e., they are scaled such that, for each *m/z* and over all the spectra, the smallest intensity value becomes zero and the largest intensity becomes one).
- Baseline subtraction (i.e., filtering the baseline signal out from raw mass spectra) is performed by linearly interpolating the local minima within variable-width windows in the spectrum.
- Peak detection and quantification aim to smooth the peak profiles in the spectra. This step is accomplished to reconstruct the theoretical Gaussian profile of the peaks. An ideal peak profile is comprised of two parts: a monotonic ascending side and a monotonic descending side. A local maximum in a spectrum is indicated as *M-peak*, that is a (*m/z*, *I*) pair in the spectrum having its intensity higher than both the previous and the next pair. A group of M-peaks for a given spectrum and the corresponding ideal reconstructed peak are depicted in Fig. 5, whereas Fig. 6 graphically shows the process of approximating the ideal Gaussian profile of any group of M-peaks.

Note that not all of the preprocessing steps can be applied to every type of spectra data. In our case, datasets *Cardiotoxicity* and *Pancreatic* came already binned, while dataset *Prostate* had altready been baseline subtracted (see subsection *Data Description* in section b), thus limiting the possibility of the preprocessing pipeline variability.



(a)

(b)

Figure 6: Peak smoothing: (a) three local M-peaks and (b) the resulting profile after smoothing

Time Series based Representation

The preprocessed mass spectrum is a sequence $S = [((m/z)_1, I_1), \ldots, ((m/z)_n, I_n)]$ such that, for each $i \in [1..n]$, $(m/z)_i$ refers to the mass-to-charge ratio whereas I_i is the associated intensity value. Since the notion of time implicitly lies in the sequence of mass-to-charge values, S can be modeled as a time series $T_S = [(x_1, t_1), \ldots, (x_n, t_n)]$, whose values x_i correspond to the intensity values I_i of S, and time steps t_i correspond to the m/z values of S. The m/z values are not involved into the proposed time series based representation of mass spectra. Thus, for the sake of simplicity, T_S is hereinafter denoted as $T = [x_1, \ldots, x_n]$.

Time series representing mass spectra are typically highly dimensional. Thus, it is desirable to model such time series with a representation which is able to reduce the dimensions while maintaining the significant variations in the time series profile. For this purpose, we employ the time series representation model called *DSA (Derivative time series Segment Approximation)* [16]. By using the DSA model, a time series of length n is transformed in linear time ($O(n)$) into a new, smaller sequence by the following main steps:

1. computation of the first derivatives of the original series to capture its significant trends (*derivation*);
2. identification of segments consisting of tight derivative points (*segmentation*);
3. representation of each segment by involving synthetic information (*segment approximation*)

Each of such steps are described next.

Derivation. The *derivation* step yields a sequence $\dot{T} = [\dot{x}_1, \ldots, \dot{x}_n]$, where $\dot{x}_i$ are first derivative estimates. We use an estimation model that is sufficiently general (i.e., independent of the underlying data distribution model) and still enough robust to outliers. This model considers for each point (except the first and the last one in the series) the slope of the line from the left neighbor to the right neighbor; neighbors are also considered when computing the derivatives of the first and last points:

$$\dot{x}_i = \begin{cases} x_{i+1} - x_i & \text{if } i = 1 \\ \frac{1}{2}(x_{i+1} - x_{i-1}) & \text{if } i \in [2..n\text{-}1] \\ x_i - x_{i-1} & \text{if } i = n. \end{cases}$$

Segmentation. The *segmentation* of a time series of length n consists in identifying a set of break-points to partition it into p ($p \ll n$) contiguous, variable-length subsequences of points (segments) having similar features. In DSA, segmentation is computed on the derivative version of a time series. Precisely, a derivative time series $\dot{T} = [\dot{x}_1, \ldots, \dot{x}_n]$ is transformed into a sequence $S_{\dot{T}} = [s_1, \ldots, s_p]$ of variable-length segments, which are defined according to a ordered list $[b_1, \ldots, b_p]$ of break-points ($b_j \in \dot{T}$, $\forall j \in [1..p]$ and $b_1 = \dot{x}_1$). More precisely, each segment s_j, $j \in [1..p]$, is the subsequence of $\dot{T}$ given by $[\dot{x}_{b_j}, \dot{x}_{b_j+1}, \ldots, \dot{x}_{b_j+k_j-1}]$, where $k_j = |s_j|$ is the size of the segment s_j (i.e., the number of points falling into s_j) and is defined as $k_j = \dot{x}_n - b_j$, if $j = p$ and $k_j = b_{j+1} - b_j$, otherwise.

The critical aspect in segmentation is to determine the segment break-points. DSA uses the sliding windows approach: a segment grows until it exceeds an error threshold, and the process repeats starting from the next point not yet considered. The key idea in the sliding windows approach to time series segmentation is to break a time series according to the first point such that the absolute difference between it and the mean of the previous points is above a certain threshold ϵ; this point becomes the anchor for the next segment to be identified in the rest of the series. Formally, let $\mu(s_j)$ denote the average of the points in a potential segment s_j of $S_{\dot{T}}$, that is $\mu(s_j) = \left(\sum_{h=0}^{k_j-1} \dot{x}_{b_j+h}\right) \Big/ k_j$, for each $j \in [1..p]$. The sequence s_j is identified as an actual segment if and only if

$$|\mu([\dot{x}_{b_j}, \ldots, \dot{x}_{b_j+h}]) - \dot{x}_{b_j+h+1}| \leq \epsilon, \ \forall h \in [0..k_j{-}2], \quad \text{and}$$

$$|\mu([\dot{x}_{b_j}, \ldots, \dot{x}_{b_j+k_j-1}]) - \dot{x}_{b_j+k_j}| > \epsilon$$

Intuitively, this condition allows for aggregating subsequent data points having very close derivatives; in such a way, the growth segment s_j represents a subsequence of points with a specific trend.

Parameter ϵ can be defined by employing simple statistics based on an index of dispersion of the (derivative) data points within the same sequence around the respective mean value. Precisely, we compute ϵ as the variance of the points in the segment s_j which is currently being identified in $\dot{T}$, i.e., $\epsilon(s_j) = \sigma^2(s_j)$. In such a way, the estimate of ϵ is tailored to the local features of the individual series.

Segment approximation. All individual segments of a derivative time series are finally modeled with a synthetic information capturing their respective main features. More precisely, each segment s_j is mapped to a pair formed by the time t_j+1, where t_j is the timestamp of the last point ($\dot{x}_{b_j+k_j-1}$) in s_j, and an angle that explains the average slope of the portion of time series bounded by s_j. This is mathematically expressed by the notion of arctangent applied to the mean of the (derivative) points in each segment.

Given a segmented derivative time series $S_{\dot{T}} = [s_1, \ldots, s_p]$, the final step of segment approximation yields a sequence $\tau = [(\alpha_1, t'_1), \ldots, (\alpha_p, t'_p)]$, called *DSA sequence*, such that

$$\begin{array}{ll} \alpha_j = \arctan(\mu(s_j)) & j \in [1..p], \\ t'_j = \begin{cases} k_j \\ t'_{j-1} + k_j \end{cases} & \begin{array}{l} j = 1 \\ j \in [2..p] \end{array} \end{array}$$

where the assumption is $t'_0 = 0$ for any DSA sequence.

It is worth pointing out that the segment approximation step of our DSA model falls into the well-known class of piecewise approximation (PA) algorithms [46]. PA approximation algorithms can be classified into three main categories: piecewise constant (PCA), which are based on polynomials having degree 0; piecewise linear (PLA), which are based on polynomials having degree 1; continuous piecewise linear, which are analogous to PLA with the additional constraint of requiring segments continuous at the knots. Piecewise linear models (basic or continuous) are typically more accurate than the piecewise constant counterpart. The segment approximation algorithm involved into the DSA model is composed by two sub-steps: computing the average of the points within the segment focusing on the derivative series, and re-mapping this average into the original time series domain by maintaining the information about the angular coefficient of the line that contains that segment. The first sub-step is equivalent to computing a PCA of the derivative series according to a least squares estimate, whereas the second sub-step aims to define piecewise linear segments from the constant ones computed in the previous step without requiring any least squares estimate. Therefore, the DSA model is able to approximate time series according to a piecewise linear approximation which is a piecewise constant approximation performed over the derivative version of the time series to be represented.

Data clustering

The proposed methodology for MS data clustering is parametric with respect to the clustering scheme; in this work, we resort to centroid-based partitional clustering [45], due to the advantages offered in terms of simplicity, execution time and space requirement. The problem of partitional clustering can be formulated as follows: given a dataset $\mathcal{D}$ and an integer $K < |\mathcal{D}|$, the goal is to divide $\mathcal{D}$ into K internally homogeneous, well-separated groups (clusters). The exemplary centroid-based partitional clustering method is the well known K-Means algorithm [45], whose outline is given as follows.

```
Select $K$ instances as the initial cluster centroids.
repeat
  Assign each instance to the closest cluster
  based on its distance to cluster centroids.
```

```
    Recompute the centroid of each cluster.
until the centroids do not change.
```

The notion of cluster centroid can be trivially given as the mean instance of that cluster, provided that all the cluster members are of equal length. Nevertheless, a more refined model to summarize sets (clusters) of DSA sequences is used here; such a model is well-suited to any centroid-based partitional clustering method (see [16]).

The similarity measure employed for comparing DSA sequences is DTW. As previously discussed, using DTW is particularly advantageous in the MS context, as it allows for reducing the implicit hardness of the crucial MS preprocessing task. Note that the learning phase of the proposed approach is based on a data clustering task and does not require a supervised classification task.

Results

The proposed time series based framework has been implemented using Java$^{\text{TM}}$programming language and tested on publicly available MS datasets to assess its effectiveness in clustering MS data.[1] In the following, we discuss the data collections of mass spectra, the way(s) to preprocess the data, the clustering algorithm along with the distance/similarity measure and the choice of number of clusters, and the quality measure(s) to numerically quantify the effectiveness of the output clusterings. The next subsection discusses the results obtained, in terms of both quantitative and qualitative evaluation.

Data description. Typical datasets in real-world MS application domains contain tens to hundreds spectra [18]. We selected dataset from the NCI's Center for Cancer Research.[2] Such dataset contain SELDI-TOF spectra obtained using different clinical studies under different mass spectrometry platforms and experimental conditions. Table 1 summarizes the main characteristics of the selected datasets, while a brief description is provided next.

Table 1: Main characteristics of test datasets

	#instances	***#attributes***	***#classes***	***size (MB)***
Cardiotoxicity	115	7,105	4	12.7
Pancreatic	181	6,771	2	18.2
Prostate	322	15,154	4	101

- Cardiotoxicity — This set contains spectra data about anthracycline-induced cardiotoxicity which has been used in a toxiproteomic analysis (see [48] for details). Spectra have been acquired in high resolution and are already binned and labeled according to four classes: definite negative (24 samples), probable negative (43 samples), probable positive (10 samples), definite positive (34 samples), where negative (resp. positive) samples refer to healthy (resp. diseased) patients;
- Pancreatic — This set contains spectra data which have been used in a study on cancer detection in presence of premalignant pancreatic cancer (see [49] for details). Spectra have been acquired in high resolution, binned and classified into two classes: control, (101 samples), pancreatic intraepithelial neoplasias (80 samples);
- Prostate — This set contains spectra data about patient samples with prostate related diseases. Spectra have been acquired in low resolution with baseline subtraction and classified into four classes: "cancer with PSA level > 10 *ng/ml*" (43 samples), "cancer with PSA level within [4..10] *ng/ml*" (26 samples), "benign with PSA level > 4 *ng/ml*" (190 samples), "no evidence of disease" (63 samples).

Preprocessing setups. As far as data preprocessing, the following sequences of operations have been performed as different *preprocessing setups*: *(S1)* baseline subtraction, peak detection, normalization; *(S2)* peak detection, baseline subtraction, normalization; *(S3)* baseline subtraction, normalization; *(S4)* peak detection, normalization; *(S5)* peak detection, baseline subtraction; *(S6)* baseline subtraction, peak detection; *(S7)* baseline subtraction; *(S8)* peak detection; *(S9)* normalization.

[1] We exploited MaSDa [47], which is a Java$^{\text{TM}}$software system that implements the functionalities described in this work.

[2] `http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`

Clustering algorithms and similarity measures. K-Means is equipped with either the Euclidean distance (L_2) or the Dynamic Time Warping (DTW); the latter was applied on the spectra represented as time series by means of the DSA time series representation model (i.e., DSA sequences), whereas the Euclidean distance is referred to as the baseline method for computing distance among MS data. Thus, Euclidean distance performs a point-to-point comparison between all the (m/z, I) pairs of any two mass spectra to be compared. On the contrary, the DTW measure performs an alignment among the numerical values that represent the segments of the corresponding DSA sequences.

Quality measures

All the datasets here involved are coupled with information about the pathological status of the patients which the spectra come from. Such an information, which is referred to as *reference classification*, is exploited to evaluate how well any clustering solution fits a predefined scheme of known classes (natural clusters), i.e., how much the clusters discovered by any method are close to the natural partition identified by the various pathological status.

Two well-established measures for assessing the quality of any given clustering solution with respect to a reference classification are used in this work, namely *F-measure* and *entropy* [51]. Given a collection $\mathcal{D}$, let $\widetilde{C} = \{\widetilde{C}_1, \ldots, \widetilde{C}_h\}$ be the desired classification of the data in $\mathcal{D}$, and $C = \{C_1, \ldots, C_K\}$ be the output partition yielded by any clustering algorithm, such that $|C_j| > 0$, $\forall j \in [1..K]$.

F-measure (F) is defined as the harmonic mean between the Information Retrieval notions of precision (P) and recall (R):

$$F = \frac{2 \times P \times R}{P + R}$$

Precision of C_j with respect to $\widetilde{C}_i$ is the fraction of the series in C_j that has been correctly classified, i.e., $P_{ij} = |C_j \cap \widetilde{C}_i|/|C_j|$. Recall of C_j with respect to $\widetilde{C}_i$ is the fraction of the series in $\widetilde{C}_i$ that has been correctly classified, i.e., $R_{ij} = |C_j \cap \widetilde{C}_i|/|\widetilde{C}_i|$. The overall precision and recall are defined as:

$$P = \frac{1}{H}\sum_{i=1}^{H} P_i \qquad R = \frac{1}{H}\sum_{i=1}^{H} R_i$$

where each P_i and R_i are equal to P_{ij^*} and R_{ij^*}, respectively, such that $j^* \in argmax_{j=1..K}\{P_{ij}, R_{ij}\}$.

In the case of entropy, for each cluster $C_j \in C$ the class distribution of data is computed as the probability $\Pr(\widetilde{C}_i|C_j)$ that an instance in C_j belongs to class $\widetilde{C}_i$. Using this class distribution, the normalized entropy of C_j is computed as

$$E_j = -\frac{1}{\log h}\sum_{i=1}^{h} \Pr(\widetilde{C}_i|C_j) \times \log\left(\Pr(\widetilde{C}_i|C_j)\right)$$

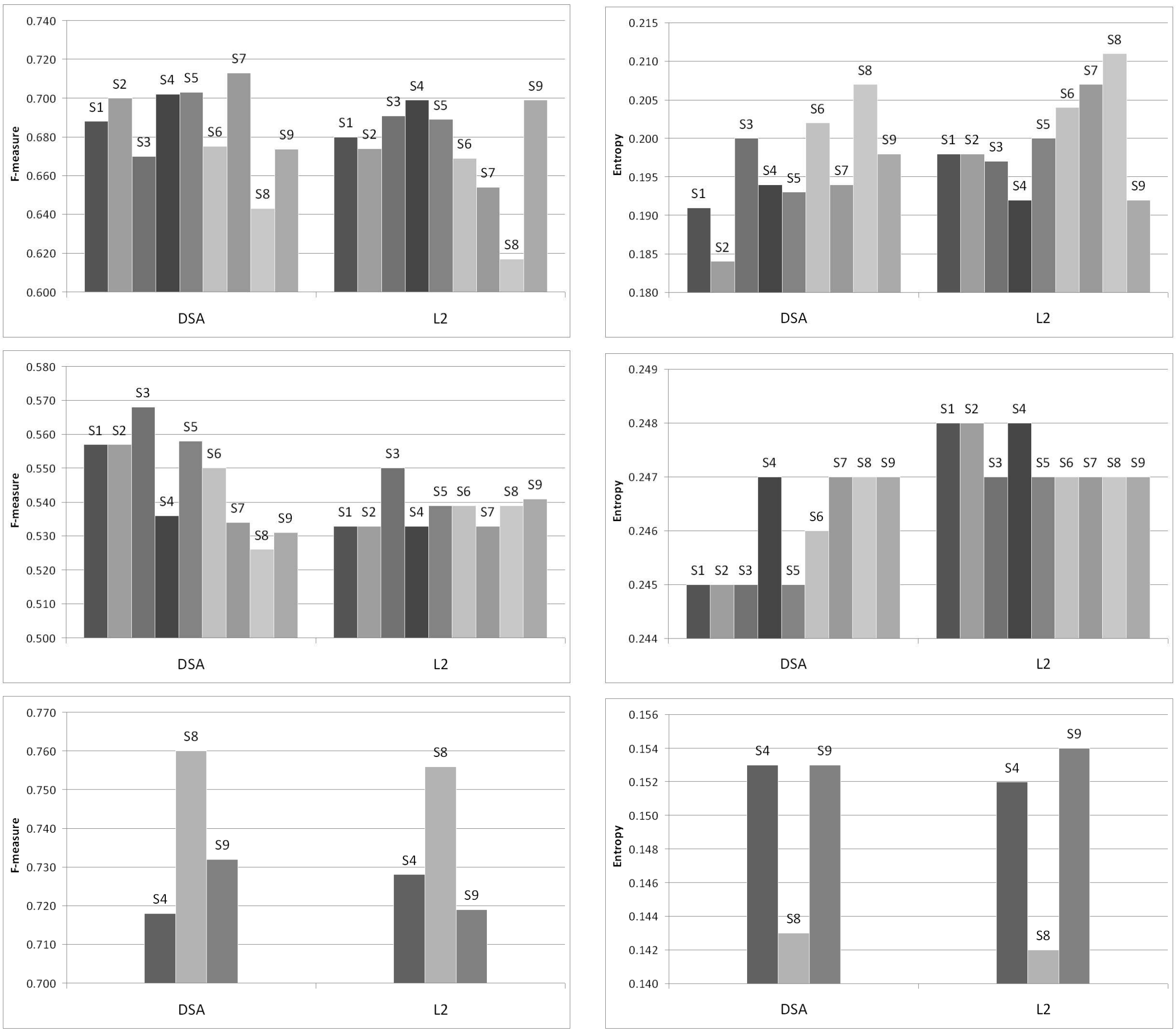
where $\Pr(\widetilde{C}_i|C_j) = |C_j \cap \widetilde{C}_i|/|C_j|$. The overall entropy ($E \in [0..1]$) is defined as the sum of the individual cluster entropies weighted by the size of each cluster:

$$E = \frac{1}{|\mathcal{D}|}\sum_{j=1}^{k} |C_j| \times E_j$$

Based on the above measures, a good clustering solution is expected to have both high F-measure and low entropy.

Discussion

For each performed test, various runs of the K-means algorithm have been performed, by varying the preprocessing setup, and compared the results obtained by using the time series based approach with the standard Euclidean approach.



(a)

(b)

Figure 7: Clustering quality results in terms of (a) F-measure, and (b) Entropy for Cardiotoxicity (top), Pancreatic (middle), and Prostate (bottom)

Quantitative evaluation. Tests have been initially focused on testing the ability of the framework to detect and distinguish all the meaningful groups in the data, that is 4 for Cardiotoxicity and Prostate and 2 for Pancreatic. Thus, for all experiments on a specific dataset, the number of clusters has been put exactly equal to the number of classes associated with that dataset.

Figure 7 shows the quality results obtained on the various datasets, and compares our DSA-based approach to the standard Euclidean distance.

The initial centroids of the K-Means algorithm were randomly chosen. For this purpose, in order to avoid for the results to be affected by random chance, all the reported accuracy values were obtained by averaging over 100 runs. A very low standard deviation (ranging between 0.001 and 0.008) was encountered on each experiment.

Clustering performances were generally affected by the selected preprocessing setup, while baseline subtraction revealed to be essential for improving the clustering quality in most cases. Note that only setups $S4$, $S8$ and $S9$ were considered for Pancreatic, since this data has been already subject to baseline subtraction. The DSA-based approach achieved reasonably good clustering results at least in Cardiotoxicity and Prostate; in Pancreatic, clustering inevitably

Table 2: Summary of clustering results on the various test datasets

	K	DSA			L_2		
		F (P/R)	E	set.	F (P/R)	E	set.
Cardiotoxicity	4	.72 (.75 / .68)	.19	$S7$	.69 (.70 / .69)	.19	$S4/9$
Cardiotoxicity	2	.76 (.78 / .73)	.38	$S2$	.67 (.67 / .67)	.45	$S4/9$
Pancreatic	2	.57 (.58 / .56)	.48	$S3$	.55 (.57 / .53)	.49	$S3$
Prostate	4	.76 (.84 / .69)	.14	$S8$	.75 (.84 / .68)	.14	$S8$
Prostate	2	.78 (.79 / .78)	.34	$S8$	.77 (.79 / .76)	.34	$S8$

Table 3: P-values for unpaired T-Test (df: 198)

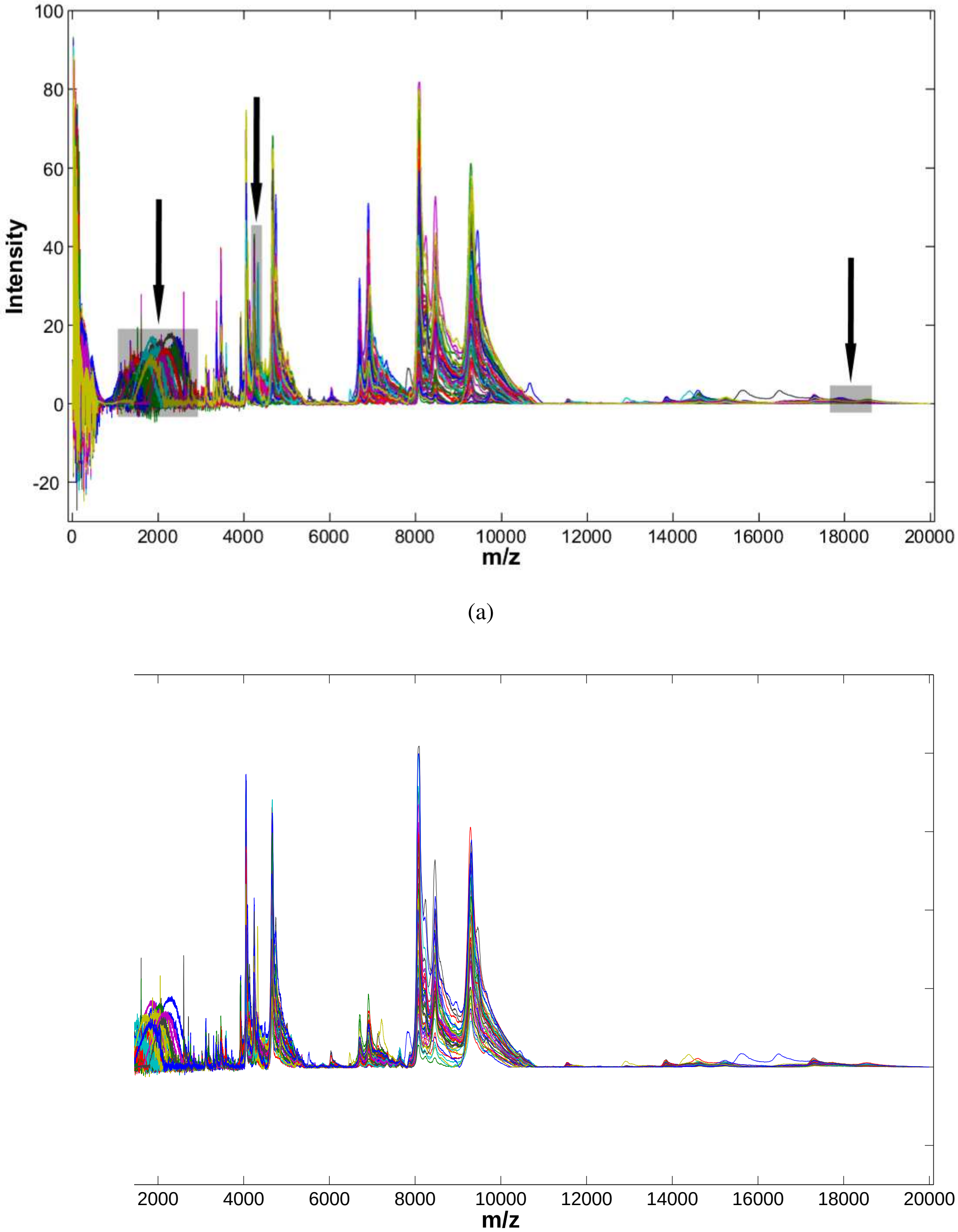
Dataset	Score	DSA vs. L_2
Cardiotoxicity (K=4)	F	5.03E-147
	E	6.03E-62
Cardiotoxicity (K=2)	F	1.83E-231
	E	1.39E-206
Pancreatic	F	4.94E-135
	E	4.88E-79
Prostate (K=4)	F	9.13E-107
	E	1.26E-79
Prostate (K=2)	F	1.47E-65
	E	1.06E-39

resulted in lower performances mainly due to the dominant presence of *m/z* values with very low intensity values and just a few characteristic trends in the spectra. Anyway, for all datasets the DSA-based approach was able to achieve higher F-measure and lower Entropy scores than the standard Euclidean approach.

The superiority of our DSA-based approach with respect to the Euclidean approach was also proved in terms of statistical significance. For this purpose, we carried out an unpaired T-Test, under the null hypothesis of no difference in the means between any two groups of performance scores (i.e., F-measure and Entropy) of the two approaches. Table 3 reports the p-values for the T-Test, where for each dataset and evaluation criterion, the 100-run pools of the DSA-based approach were compared to those obtained by the Euclidean approach. As we can observe in the table, the p-values are extremely low, which allows us to reject the null hypothesis, at $\alpha = 0.01$ significance level, in all the cases.

Table 2 summarizes the quality results (i.e. F-measure along with corresponding precision and recall, and Entropy) referring to the best preprocessing setups; for each dataset and method, the best preprocessing setup is that leading to the highest quality in terms of F-measure. This table also includes results obtained by a two-class task of clustering; more precisely, for Cardiotoxicity and Prostate, the data assigned with the definite cancer (diseased) or the definite non-cancer (healthy) classes are selected; then clustering is performed only on the selected subsets, aiming to give emphasis on distinguishing solely the extreme classes.

Performing K-Means with dynamic time warping on DSA sequences behaved as good as or better than standard Euclidean distance on the original spectra, up to a 10% improvement (Cardiotoxicity, 2 classes). It should be noted that the advantage of using the dynamic time warping on DSA sequences becomes important since the DSA model yields compact yet dense representations of the original spectra. The lower dimensionality of the spectra-series achieved by DSA is beneficial for the efficiency of the clustering task (and further post-processing analysis), while not affecting

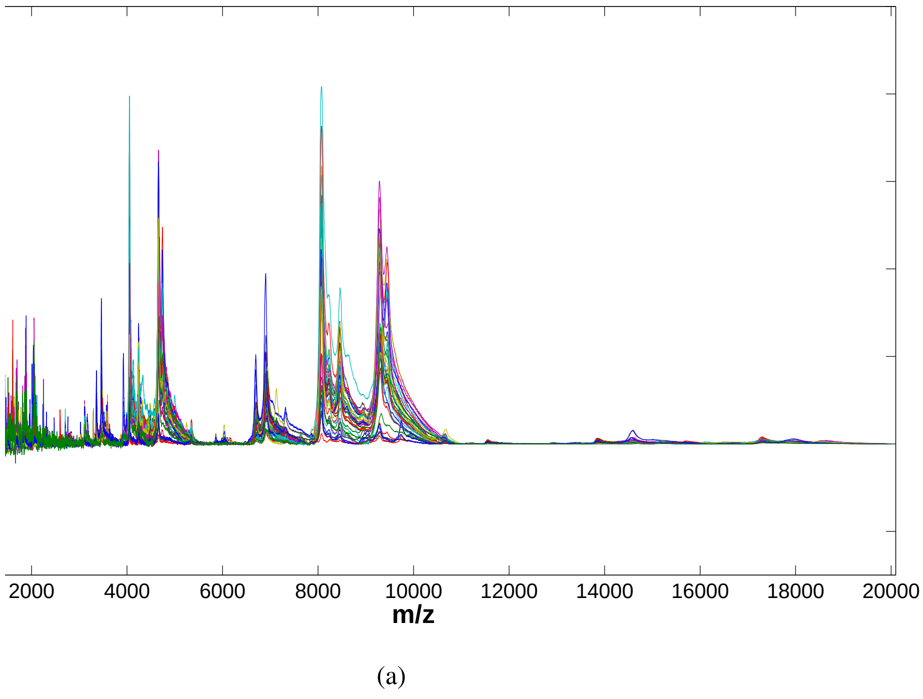


(a)

(b)

Figure 8: Clusters vs. natural classes from Prostate: (a) cluster and (b) class of cancer with PSA>10 ng/ml. Arrows highlight spectra portions that characterize cancer condition.

negatively the clustering effectiveness. To report some details, the following data compression ratios were achieved



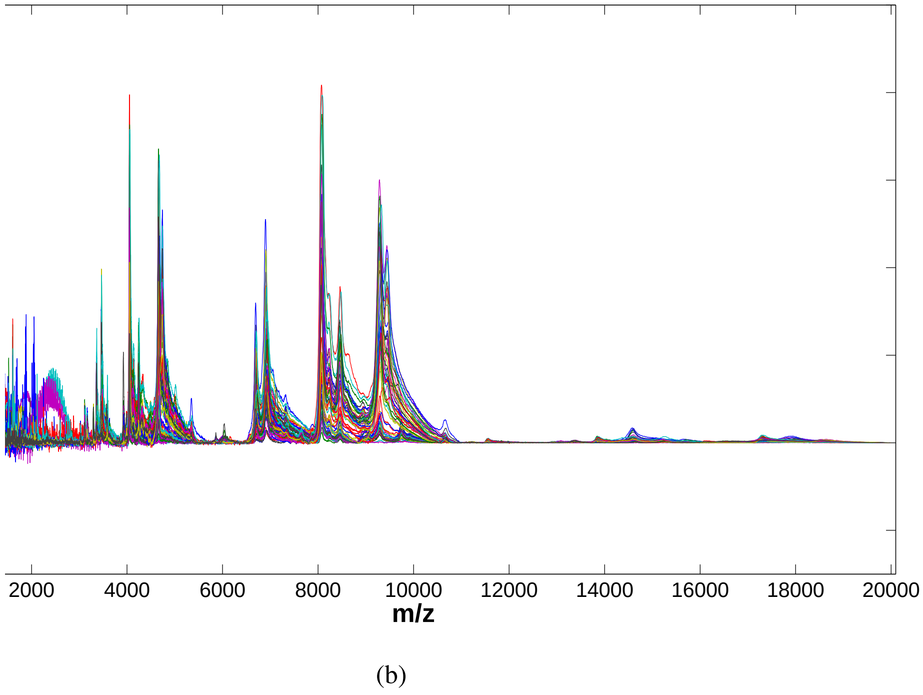
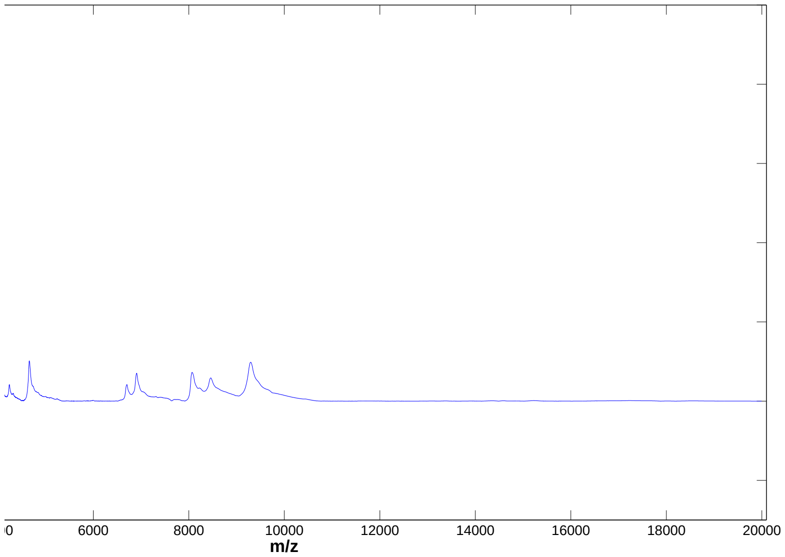
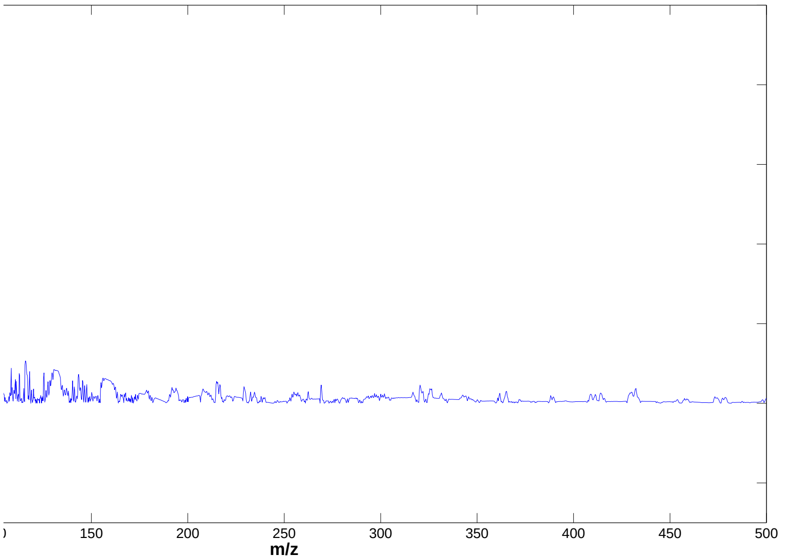


Figure 9: Clusters vs. natural classes from Prostate: (a) cluster and (b) class of no evidence of disease.

(on the preprocessed spectra): 64% on Cardiotoxicity, 65% on Pancreatic, and 97% on Prostate. The latter result is particularly significant since it shows that a very high compression was obtained for a dataset on which DSA still



(a)

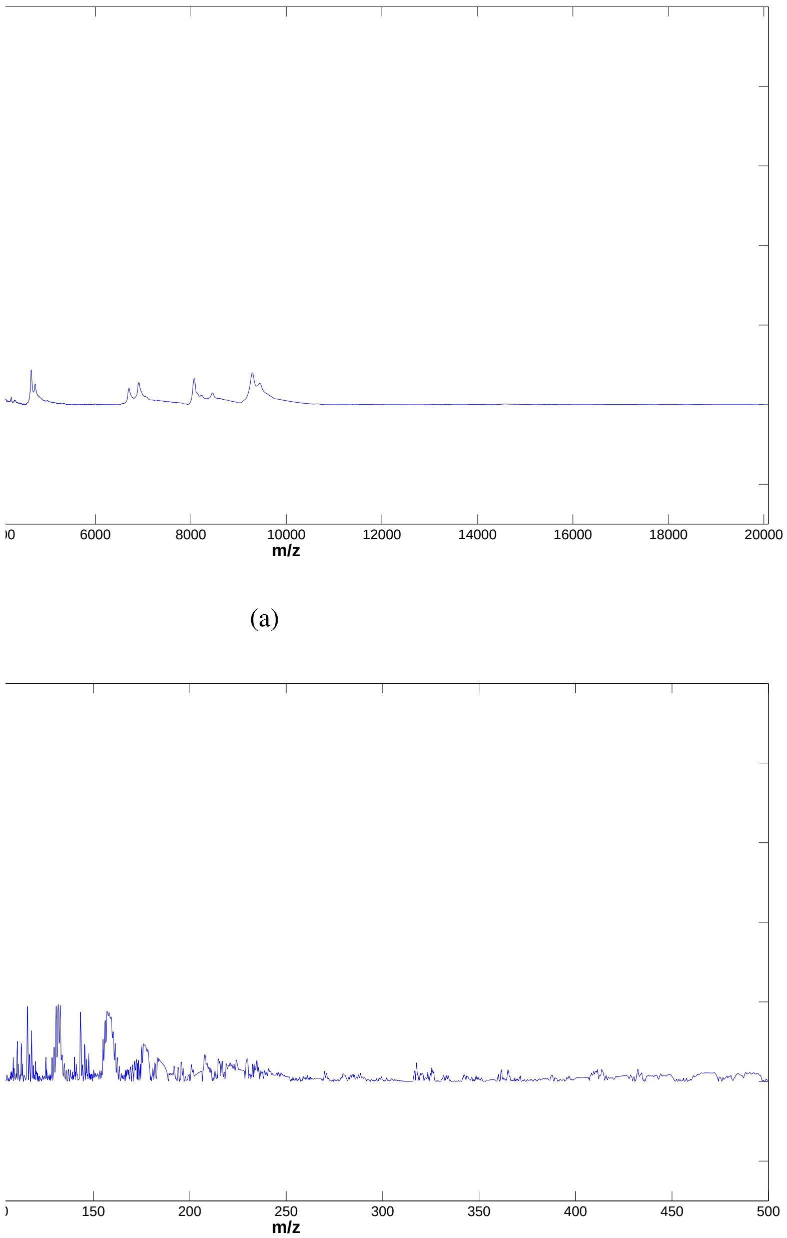


(b)

Figure 10: Prostate with PSA>10: point-to-point average absolute difference between the spectra in the cluster and the spectra in the corresponding natural class within the m/z ranges of $[0..20,000]$ (a) and $[0..500]$ (b)

performs very closely to (slightly better than) the standard approach based on L_2.

Qualitative evaluation. It is useful to gain an insight into the clusters detected by the proposed method with respect to the natural classes, i.e., the ideal classifications provided along with the selected datasets. For the sake of brevity of the presentation, only the Prostate dataset is taken into account here — however, similar considerations hold for the



(a)

(b)

Figure 11: Prostate no evidence of disease: point-to-point average absolute difference between the spectra in the cluster and the spectra in the corresponding natural class within the m/z ranges of $[0..20,000]$ (a) and $[0..500]$ (b)

remaining datasets.

Figures 8 and 9[3] refer to the "cancer with PSA>10 ng/ml" and "no evidence of disease" natural classes of the Prostate dataset, respectively (cf. Section Results). Both the figures report on the spectra of the detected cluster

[3] High-quality images for these figures are available at http://polifemo.deis.unical.it/~gtradigo/jocs2010/

(on the top) vs. the spectra belonging to the corresponding natural class (on the bottom). The main goal of this analysis is to provide a graphical intuition of how well the clusters detected look similar to the corresponding natural classes. This can be easily evinced by noting that Figure 8(a) is highly similar to Figure 8(b) (for the ‘cancer with with PSA>10 ng/ml” class), as well as Figure 9(a) is very close to Figure 9(b) (for the “no evidence of disease” class), thus suggesting that most spectra have been correctly recognized.

Furthermore, the authors of the study in [50] have discovered a number of main patterns that allow for discriminating between the “cancer with with PSA>10 ng/ml” (diseased individuals) and the “no evidence of disease” class (healthy individuals) of the Prostate dataset. These patterns (which are highlighted in Figure 8(a)) refer to portions of the *m/z* axis where the spectra of the diseased individuals significantly differ from the spectra of the healthy individuals. The discriminating power of these patterns has been kept nearly intact in the clusters detected by the proposed method. Indeed, like the natural classes, the clusters of diseased (Figure 8(a)) and healthy (Figure 9(c)) individuals mainly differ from each other right in the portions of the spectrum that contain the discriminating patterns, thus evidencing that the proposed method has allowed for correctly maintain the information about the main biomarkers responsible of distinguishing among pathological states.

Figures 10 and 11 provide more precise information about the closeness of the clusters in Figures 8(a) and 9(a) to the corresponding natural classes depicted in Figures 8(b) and 9(b), respectively. In particular, Figures 10(a)-(b) illustrate the point-to-point average absolute difference between the intensity values of the spectra from the cluster of diseased individuals (Figure 8(a)) and the intensity values of the spectra from the corresponding natural class (Figure 8(b)); Figure 10(a) shows all the values in the m/z axis, whereas 10(b) focuses only the [0..500] m/z range, which is the range where the spectra intensity values have the broadest variation. Similarly, Figures 10(a)-(b) refer to the cluster/class of healthy individuals (Figures 9(a)-(b)). Figures 10 and 11 clearly show that the average distance between clusters and classes is very low, even in the m/z range between 0 and 500. Indeed, such a difference is always lower than 20, while being lower than 10 in the majority of the m/z values.

Conclusion

We presented a data mining framework for MS data, which aims to cluster spectra modeled as time series. In this respect, we proposed a time series based representation model for MS data which produces a high dimensionality reduction while preserving and emphasizing the main features. Moreover, such a representation revealed to be particularly suitable to capture some aspects in the spectra that the preprocessing phase has not been able to correctly identify. Tests have been performed on publicly available proteomic datasets, and has shown that the proposed approach provides a useful support for the identification of discriminant biomarkers which experts may associate to pathological states. This work offers new interesting topics to be investigated as evaluation of effects of dynamically changing pipelines, as well as using of additional methods instead of using k-means.

References

[1] R. Aebersold, M. Mann, Mass spectrometry-based proteomics, Nature 422 (2003) 198–207.

[2] M. Karas and F. Hillenkamp, Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons, Analytical Chemistry 60 (1988) 259–280.

[3] T.W. Hutchens and T.T. Yip, New desorption strategies for the mass spectrometric analysis of macromolecules, Rapid Communications in Mass Spectrometry 7(7) (1993) 576–580.

[4] F. Ahmed, Application of maldi/seldi mass spectrometry to cancer biomarker discovery and validation, Current Proteomics 5 (2008) 224–252.

[5] G. Chen, B. N. Pramanik, Lc-ms for protein characterization: current capabilities and future trends, Expert Review of Proteomics 5 (2008) 435–44.

[6] F. Ahmed, The role of capillary electrophoresis-mass spectrometry to proteome analysis and biomarker discovery, Journal of Chromatography B 877(22) (2009) 1963–81.

[7] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, Comparison of Statistical Methods for Classification of Ovarian Cancer Using Mass Spectrometry Data, Bioinformatics 21(13) (2003) 1636–1643.

[8] M. Wagner, D. Naik, and A. Pothen, Protocols for Disease Classification from Mass Spectrometry Data, Proteomics 3(9) (2003) 1692–1698.

[9] P. Geurts, M. Fillet, D. de Seny, M.A. Meuwis, M. Malaise, M.P. Merville, and L. Wehenkel, Proteomic mass spectra classification using decision tree based ensemble methods, Bioinformatics 21(14) (2005) 3138–3145.

[10] J. Prados, A. Kalousis, J.C. Sanchez, L. Allard, O. Carrette, and M. Hilario, Mining mass-spectra for diagnosis and biomarker discovery of cerebral accidents, Proteomics 4(6) (2004) 2320–2332.

[11] S. Meleth, I.-E. Eltoum, L. Zhu, D. Oelschlager, C. Piyathilake, D. Chhieng, and W.E. Grizzle, Novel approaches to smoothing and comparing SELDI TOF spectra, Cancer Informatics 1(1) (2005) 78–85.

[12] M. Wagner, D.N. Naik, A. Pothen, S. Kasukurti, R.R. Devineni, B.L. Adam, O.J. Semmes, and G.L. Wright, Computational protein biomarker prediction: a case study for prostate cancer, BMC Bioinformatics 5 (2004) 26.

[13] K.A. Baggerly, J.S. Morris, J. Wang, D. Gold, L.C. Xiao, and K.R. Coombes, A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples, Proteomics 3 (2003) 1667–1672.

[14] E.F. Petricoin 3rd, A.M. Ardekani, B.A. Hitt, P. Levine, V.A. Fusaro, and S. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, and L.A. Liotta, Use of proteomic patterns in serum to identify ovarian cancer, Lancet 359 (2002) 572–577.

[15] V. A. Emanuele, B. M. Gurbaxani, Benchmarking currently available seldi-tof ms preprocessing techniques, Proteomics 9(7) (2009) 1754–1762.

[16] F. Gullo, G. Ponti, A. Tagarelli, S. Greco, A Time Series Representation Model for Accurate and Fast Similarity Detection, Pattern Recognition 42 (11) (2009) 2998–3014.

[17] J.M. Sorace and M. Zhan, A data review and re-assessment of ovarian cancer serum proteomic profiling, BMC Bioinformatics 4 (2003) 24.

[18] K.R. Coombes, K.A. Baggerly, and J.S. Morris, Pre-Processing Mass Spectrometry Data, Fundamentals of Data Mining in Genomics and Proteomics, Kluwer, Boston, 2007.

[19] B. Williams, S. Cornett, B.M. Dawant, A. Crecelius, B. Bodenheimer, and R. Caprioli, An algorithm for baseline correction of MALDI mass spectra, in: Proc. ACM Southeast Regional Conf., 2005, pp. 137–142.

[20] A.C. Sauve and T.P. Speed, Normalization, Baseline Correction and Alignment of High-Throughput Mass Spectrometry Data, in: Proc. Genomic Signal Processing and Statistics Conference, 2004.

[21] 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 (1999) 179–193.

[22] W.E. Wallace, A.J. Kearsley, and C.M. Guttman, An Automated Peak Identification/Calibration Procedure for High-Dimensional Protein Measures From Mass Spectrometers, Analytical Chemistry 76(9) (2004) 2446–2452.

[23] Y. Yasui, D. McLerran, B.L. Adam, M. Winget, M. Thornquist, and Z. Feng, An Operator-Independent Approach to Mass Spectral Peak Identification and Integration, Journal of Biomedicine and Biotechnology 4 (2003) 242–248.

[24] J.W.H. Wong, G. Cagney, and H.M. Cartwright, SpecAlign - processing and alignment of mass spectra datasets, Bioinformatics 21(9) (2005) 2088–2090.

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

[26] S.M. Carlson, A. Najmi, and H.J. Cohen, Biomarker clustering to address correlations in proteomic data, Proteomics 7(7) (2007) 1037–46.

[27] E. Lange, C. Gropl, O. Schulz-Trieglaff, A. Leinenbach, C. Huber, and K. Reinert, A geometric approach for the alignment of liquid chromatography-mass spectrometry data, Bioinformatics 23 (2007) 1273–81.

[28] F. Gullo, G. Ponti, A. Tagarelli, G. Tradigo, P. Veltri, MSPtool: A Versatile Tool for Mass Spectrometry Data Preprocessing, in: Proc. IEEE Int. Symposium on Computer-Based Medical Systems (CBMS), 2008, pp. 209–214.

[29] N. Barbarini, P. Magni, Accurate peak list extraction from proteomic mass spectra for identification and profiling studies, BMC Bioinformatics 11 (1) (2010) 518.

[30] V. A. Emanuele, B. M. Gurbaxani, Quadratic variance models for adaptively preprocessing seldi-tof mass spectrometry data, BMC Bioinformatics 11 (2010) 512.

[31] S. Chen, M. Li, D. Hong, D. Billheimer, H. Li, B. J. Xu, Y. Shyr, A novel comprehensive wave-form ms data processing method, Bioinformatics 25(6) (2009) 808–814.

[32] M.K. Titulaer, I. Siccama, L.J. Dekker, A.L.C.T. van Rijswijk, R.M.A. Heeren, P.A. Sillevis Smitt, and T.M. Luider, A database application for pre-processing, storage and comparison of mass spectra derived from patients and controls, BMC Bioinformatics 7 (2006) 403.

[33] J. Hartler, G.G. Thallinger, G. Stocker, A. Sturn, T.R. Burkard, E. Krner, R. Rader, A. Schmidt, K. Mechtler, and Z. Trajanoski, Maspectras: a platform for management and analysis of proteomics lc-ms/ms data, BMC Bioinformatics 8 (2007) 197.

[34] M. Cannataro, P. Veltri, MS-Analyzer: preprocessing and data mining services for proteomics applications on the Grid, Concurrency and Computation: Practice and Experience 19(15) (2007) 2047–2066.

[35] A. Biswas, K. C. Mynampati, S. Umashankar, S. Reuben, G. Parab, R. Rao, V. S. Kannan, S. Swarup, MetDAT: a modular and workflow-based free online pipeline for mass spectrometry data processing, analysis and interpretation, Bioinformatics 26 (20) (2010) 2639–2640.

[36] D.W. Powell, C.M. Weaver, J.L. Jennings, K.J. McAfee, Y. He, P.A. Weil, and A.J. Link, Cluster Analysis of Mass Spectrometry Data Reveals a Novel Component of SAGA, Molecular and Cellular Biology 24(16) (2004) 7249–7259.

[37] D.P. De Souza, E.C. Saunders, M.J. McConville, and V.A. Liki, Progressive peak clustering in GC-MS Metabolomic experiments applied to Leishmania parasites, Bioinformatics 22(11) (2006) 1391–1396.

[38] D.J. Slotta, L.S. Heath, N. Ramakrishnan, R. Helm, and M. Potts, Clustering mass spectrometry data using order statistics, Proteomics 3 (2003) 1687–1691.

[39] H. Zheng, S.S. Anand, J.G. Hughes, and N.D. Black, Methods for Clustering Mass Spectrometry Data in Drug Development, in: Proc. Workshop on Intelligent Data Analysis in Medicine and Pharmacology, 2000.

[40] H. Bensmail, J. Golek, M.M. Moody, J.O. Semmes, and A. Haoudi, A novel approach for clustering proteomics data using Bayesian fast Fourier transform, Bioinformatics 21(10) (2005) 2210–2224.

[41] K. Flikka, J. Meukens, K. Helsens, J. Vandekerckhove, and I. Eidhammer and K. Gevaert and L. Martens, Implementation and application of a versatile clustering tool for tandem mass spectrometry data, Proteomics 7(18) (2007) 3245–58.

[42] J. Dutkowski and A. Gambin, On consensus biomarker selection, BMC Bioinformatics 8 (Suppl 5) (2007) S5.

[43] T. W. Liao, Clustering of Time Series Data—A Survey, Pattern Recognition 38 (2005) 1857–1874.

[44] D. J. Berndt and J. Clifford, Using Dynamic Time Warping To Find Patterns in Time Series, in: Proc. AAAI Workshop on Knowledge Discovery in Databases, 1994, pp. 359–370.

[45] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, Prentice-Hall, 1988.

[46] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition),

Springer, 2009.

[47] F. Gullo, G. Ponti, A. Tagarelli, G. Tradigo, P. Veltri, MaSDA: A System for Analyzing Mass Spectrometry Data, Computer Methods and Programs in Biomedicine (CMPB) 95 (2) (2009) S12–S21.

[48] E.F. Petricoin 3rd, V. Rajapaske, E.H. Herman, A.M. Arekani, S. Ross, D. Johann, A. Knapton, J. Zhang, B.A. Hitt, T.P. Conrads, T.D. Veenstra, L.A. Liotta, and F.D. Sistare, Toxicoproteomics: serum proteomic pattern diagnostics for early detection of drug induced cardiac toxicities and cardioprotection, Toxicologic Pathology 32 Suppl 1 (2004) 122–130.

[49] S.R. Hingorani, E.F. Petricoin 3rd, A. Maitra, V. Rajapakse, C. King, M.A. Jacobetz, S. Ross, T.P. Conrads, T.D. Veenstra, B.A. Hitt, Y. Kawaguchi, D. Johann, L.A. Liotta, H.C. Crawford, M.E. Putt, T. Jacks, C.V. Wright, R.H. Hruban, A.M. Lowy, and D.A. Tuveson, Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse, Cancer Cell 4(6) (2003) 437–450.

[50] E.F. Petricoin 3rd, D.K. Ornstein, C.P. Paweletz, A. Ardekani, P.S. Hackett, B.A. Hitt, A. Velassco, C. Trucco, L. Wiegand, K. Wood, C.B. Simone, P.J. Levine, W.M. Linehan, M.R. Emmert-Buck, S.M. Steinberg, E.C. Kohn, and L.A. Liotta, Serum Proteomic Patterns for Detection of Prostate Cancer, Journal of the National Cancer Institute 94(20) (2002) 1576–1578.

[51] C.J. van Rijsbergen, Information Retrieval, Butterworths, 1979.