Expert Review of Proteomics

ISSN: 1478-9450 (Print) 1744-8387 (Online) Journal homepage: http://www.tandfonline.com/loi/ieru20

Machine learning approaches in MALDI-MSI: clinical applications

Manuel Galli, Italo Zoppis, Andrew Smith, Fulvio Magni & Giancarlo Mauri

To cite this article: Manuel Galli, Italo Zoppis, Andrew Smith, Fulvio Magni & Giancarlo Mauri (2016): Machine learning approaches in MALDI-MSI: clinical applications, Expert Review of Proteomics, DOI: 10.1080/14789450.2016.1200470

To link to this article: http://dx.doi.org/10.1080/14789450.2016.1200470

View supplementary material

Accepted author version posted online: 20 Jun 2016.
Published online: 23 Jun 2016.

Submit your article to this journal

Article views: 2

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=ieru20

Download by: [RMIT University Library] **Date:** 25 June 2016, At: 07:39

EXPERT REVIEW OF PROTEOMICS, 2016
http://dx.doi.org/10.1080/14789450.2016.1200470

Taylor & Francis
Taylor & Francis Group

REVIEW

Machine learning approaches in MALDI-MSI: clinical applications

Manuel Galli [a]*, Italo Zoppis [b]*, Andrew Smith[a], Fulvio Magni [a] and Giancarlo Mauri [b]

[a]Department of Medicine and Surgery, University of Milano Bicocca, Monza Brianza, Italy; [b]Department of Informatics, Systems and Communication, University of Milano Bicocca, Milano, Italy

ABSTRACT

Introduction: Despite the unquestionable advantages of Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry Imaging in visualizing the spatial distribution and the relative abundance of biomolecules directly *on-tissue*, the yielded data is complex and high dimensional. Therefore, analysis and interpretation of this huge amount of information is mathematically, statistically and computationally challenging.

Areas covered: This article reviews some of the challenges in data elaboration with particular emphasis on *machine learning* techniques employed in clinical applications, and can be useful in general as an entry point for those who want to study the computational aspects. Several characteristics of data processing are described, enlightening advantages and disadvantages. Different approaches for data elaboration focused on clinical applications are also provided. Practical tutorial based upon Orange Canvas and Weka software is included, helping familiarization with the data processing.

Expert commentary: Recently, MALDI-MSI has gained considerable attention and has been employed for research and diagnostic purposes, with successful results. Data dimensionality constitutes an important issue and statistical methods for information-preserving data reduction represent one of the most challenging aspects. The most common data reduction methods are characterized by collecting independent observations into a single table. However, the incorporation of relational information can improve the discriminatory capability of the data.

ARTICLE HISTORY
Received 2 May 2016
Accepted 8 June 2016
Published online
23 June 2016

KEYWORDS
Mass spectrometry imaging; machine learning; classification; feature selection; clustering; MALDI

Downloaded by [RMIT University Library] at 07:39 25 June 2016

1. Introduction

Matrix-Assisted Laser Desorption/Ionization – Mass Spectrometry Imaging (MALDI-MSI) is a powerful technology that allows the evaluation of the spatial distribution and relative abundance of biomolecules directly *on-tissue* [1,2], without the need of any labeling or extraction processes that could compromise the molecular structure and mask the presence of altered expression of the analytes of interest, i.e. when these alterations are present in a small area of the tissue. Moreover, the fact that MALDI is capable of ionizing a widespread range of molecules makes it suitable for explorative research, since it does not require any prior knowledge regarding the chemical nature of the molecules to be investigated. For these reasons, MALDI-MSI has been widely employed in several fields with successful results, from oncology and immunology to forensics and pharmacology [3–8].

Despite all the unquestionable advantages of MALDI-MSI, the yielded data results in being complex and high dimensional, in terms of amount of information and features to be extracted, even from a single tissue slice. Therefore, computational analysis of MSI data and mining procedures are challenging to be met [9,10].

The dimensionality of the data is strictly dependent on the spatial resolution and the mass resolution: the former is related to the capability of detecting small features in the examined tissue section, but requires a higher number of mass spectra to be acquired by lowering the distance between two consecutive pixels; a high mass resolution, on the other hand, allows for a better peak resolution, thus for a better identification of putative biomarkers by providing more accurate mass values, but increasing the sample rate (namely, the number of data points per spectrum) leads to higher file sizes, more challenging in terms of storage and computational purposes [11,12].

Moreover, another promising strategy could be the integration of proteomic data with other different sources of information (such as genetics and genomics, metabolomics, and histology), or even the use of relationships built on proteomic profiles to predict the disease membership group of some patients, in particular classification problems, as it has been proved by applying an efficient inferential strategy for genomics [13–15].

One of the aims of this article is to provide the reader with a brief overview onto the way in which the MSI data can be processed and elaborated, with particular attention to biological translatability. Many pieces of software have been employed for the purposes described throughout the article, and they comprise both software that requires programming and software that is ready to be used.

Among the former, Matlab (http://uk.mathworks.com/products/matlab) [16] and R (https://www.r-project.org/) [17] are

CONTACT Manuel Galli m.galli27@campus.unimib.it Via Cadore 48, 20900 Monza Brianza, Italy
*These authors contributed equally to this work.

Supplemental data for this article can be accessed here.

© 2016 Informa UK Limited, trading as Taylor & Francis Group

Downloaded by [RMIT University Library] at 07:39 25 June 2016

the most commonly used, since custom scripts, along with the presence of lots of additional packages, can provide the ability to achieve potentially every aim. This, in turn, guarantees the possibility to tweak the analysis by editing every parameter, combining different approaches, and so on. The fact that the software requires the knowledge of the programming language, however, makes its usage harder, and steps of quality check have to be performed to assess the reliability of the script.

Commercial software, such as ClinProTools (https://www.bruker.com/service/support-upgrades/software-downloads/mass-spectrometry.html) [18] or SCiLS Lab [19], on the other hand, can provide a very clean and intuitive user interface and perform analyses of high quality, but they can often only process data produced by instruments of that particular brand and they lack customizability. The presence of a variety of software for statistical analysis can extend the power of a MALDI-MSI analysis, by yielding more robust and reliable results.

In order to be able to employ other software, spectral files and/or peak list matrix files must be exported in a more common file format, to be imported into pieces of software of more general use. Mass spectra can be exported more commonly as imzML files [20], while peak list matrices can be exported as comma separated values files.

Orange Canvas and Weka open-source software will be considered in this article.

Orange Canvas [21,22] aims to be simple to use, highly customizable, and versatile at the same time: a graphical user interface makes it easy to perform statistical analyses, while preserving the capability to tweak the analysis by setting many parameters and allowing the scripting through Python programming (http://orange.biolab.si/toolbox/) [23].

Weka [24,25] is another example of free and open-source software aimed at performing statistical analysis on data matrices, focused on *machine learning* applications. Like Orange Canvas, it can be expanded with custom scripts written in Java, allowing for a more customizable usage.

Throughout this article, in the 'Tutorial' sub-sections (3.1.2, 3.2.2, and 3.3.2), an example workflow is presented, in order to provide the reader with a starting point for a statistical analysis of MALDI-MSI data with the aforementioned software.

Among other pieces of software that can be used to perform statistical analysis of high quality, Rapid Miner [26] can be cataloged as a software with a clean and intuitive user interface allowing extensive tweaking of algorithm parameters and analysis at the same time. It has in fact been employed in several applications using mass spectrometric data [27,28].

In this article, we will account for three main points. (1) At first, the structure of the data obtained from a MALDI-MSI analysis is explained, in order to make the reader aware of the needs and problematics related to the data and its processing, which is then presented more in detail. (2) Once the data has been processed to guarantee reproducibility and avoid artifacts, the data mining and elaboration phase is described by highlighting three of the most common processes for solving clinical problems: clustering, feature selection, and classification. (3) For each process, the basic statistic concepts are provided, along with examples of applications in the clinical practice and a tutorial to achieve the proposed aims via Orange Canvas and Weka.

2. MALDI-MSI data

2.1. Data preprocessing

Raw data collected after a MALDI-MSI analysis is de facto made of individual and independent spectra, which are generally unaligned and noisy, due to several factors related with the electronic nature of the instrument, sample heterogeneity, and sample preparation. In fact, the instrument does not perform the same way through time, and sample preparation and type can affect the quality of the obtained data [29]. This leads to fluctuations in the measured masses and in the *in situ* extraction of the analytes that could generate artifacts, hindering the discovery process. Spectral preprocessing is aimed at reducing both technical and analytical variability or artifacts, thus allowing fair comparisons among spectra acquired within the same analysis and in distinct analyses, in order to provide a more reliable elaboration of the data [30,31].

2.1.1. Smoothing

As previously mentioned, raw spectra present a quite consistent amount of noise, consisting of electrical background signal of the instrument itself and chemical noise coming from impurities in the sample. The shape of the peaks is therefore altered and *peak picking* algorithms struggle to define peaks out of the noise in the feature extraction phase [31]. The smoothing process discards the fluctuations in the spectrum profile related to the noise, allowing for a more reliable *peak picking* after yielding more defined peaks. This is a critical step, since aggressive smoothing can lead to information loss due to the possible removal of low-intensity signals or unresolved peaks.

The most employed smoothing algorithm is the *Savitzky-Golay filter*, which is able to fully preserve the intensity and the width of peaks. The filter fits subsets of consecutive data points with a low-degree polynomial function using the *linear least squares* method in order to straighten the noisy line of the spectrum [32].

2.1.2. Baseline subtraction

The baseline of a spectrum is the line connecting points with lowest intensities on which the entire spectrum lays. The baseline is again made of electrical and chemical background, which in turn hinders the feature extraction process (*peak picking*) by altering the peak intensities. The baseline subtraction process brings the spectrum onto the *x*-axis, for a more reproducible *peak picking* [31].

The *TopHat* algorithm uses the morphological operations of erosion and dilatation to remove the baseline of a spectrum. The *iterative convolution* algorithm, on the other hand, iteratively fits a polynomial function in such a way that, for each iteration, the values of all the data points that are above the polynomial are replaced with the value of the polynomial itself; the algorithm stops when the change between two consecutive iterations is smaller than a chosen threshold or when the set number of iterations is reached [33].

2.1.3. Normalization

Normalization is the process that multiplies all the intensity values in the mass spectrum by a *scaling factor* $(1/f)$, resulting

Downloaded by [RMIT University Library] at 07:39 25 June 2016

in a broadening or narrowing of the intensity axis. This ensures reproducible comparisons among spectra by adjusting the intensity axis to a common scale. It is a crucial step, since it can introduce artifacts that mislead the interpretation of the results [34].

$$f = \left(\sum_{i=1}^{n} |y|_i^p \right)^{\frac{1}{p}}$$

Before performing normalization, a transformation of the data might be necessary in order to flatten the differences in the variance of all the peak intensities and to make the data homoscedastic and normally distributed. Square root and logarithm of the peak intensities have been proposed for achieving this aim [35].

The most employed normalization algorithm is the *total ion count* (TIC) method (a *p*-norm with $p = 1$): all the intensities of each spectrum in the dataset are divided by the spectrum total current (i.e. the sum of all the intensities), in such a way that each spectrum has the same integrated area under the curve (equal to 1). This method is more suitable when comparing spectra with similar number of signals but can introduce artifacts if there is a compound that is much more present than the others (such as insulin in pancreas), since the TIC normalization would result in the suppression of all the other signals in this case. To overcome these limitations, the normalization can be performed either excluding the most intense peak(s) from the TIC or using only the most intense peak(s) as TIC [34].

The *root mean square* algorithm divides the spectrum by the square root of the sum of the intensity values squared. It is again based upon the assumption that the intensities of all peaks across the dataset are quite similar, thus it can suffer from artifacts generated by the presence of intense signals [34].

The *median* normalization divides each spectrum by the median of the intensity values in the spectrum. This method has been found to be more robust to different spectral preprocessing methods and to different peak intensities, and to suppress the artifacts generated by high-intensity peaks [34].

2.1.4. Peak picking and alignment

Once the preprocessing has ensured that the data has been purified from analytical variability coming from the sample content and the instrument's nature, the *peak picking* extracts the features that characterize the spectrum, namely, the list of the *mass-to-charge* (*m/z*) values of the signals along with their intensities or areas under the curve. This feature extraction process leads to a consequent reduction of the data that makes algorithms computationally faster and more efficient.

The majority of the algorithms employed for this task make use of a function to estimate the noise (e.g. the median of the absolute deviation of points in a window) in order to choose only the local maxima with a *signal-to-noise* ratio over a certain threshold that come out from the noise [36]. However, this approach is prone to generate false positives, due to the difficulty in discriminating the signals from the noise and to some differences in the baseline across the spectrum [36].

In order to overcome possible artifacts coming from picking false positive peaks which belong to the noise, new methods (such as the Orthogonal Matching Pursuit (OMP)) have been developed in order to evaluate the shape of a peak (through a mathematical function) rather than its intensity. The OMP algorithm models the peaks as shape functions (e.g. Gaussian curves), with a high level of robustness to variations in the peak shape and symmetry [9].

In order to prevent slightly analytical variations in the *m/z* values from being seen as distinct peaks, all peak values must be aligned (to a reference list or to each other), to ensure more consistent and coherent results by selecting the exact same peak across the dataset. The best way to achieve this is to align the peaks to a reference list of peaks, that can be constituted, e.g., by the peaklist of the average spectrum of the dataset [37].

2.2. The data cube

An MSI dataset is structured as a 'data cube' (Figure 1), which is the result of the acquisition of one mass spectrum for each pixel of a digitalized tissue image. Therefore, for each spatial coordinate, the presence and the relative amount of biomolecules are recorded by the mass spectrum itself. On the other hand, when considering a *m/z* value of interest, the spatial distribution of the corresponding compound (with that specific *m/z*) can be displayed by coloring each pixel according to the intensity (i.e. relative abundance) of that *m/z* value in the related spectrum. Putative regions of interest can be highlighted by a specific localization of the selected analyte(s) *on-tissue*.

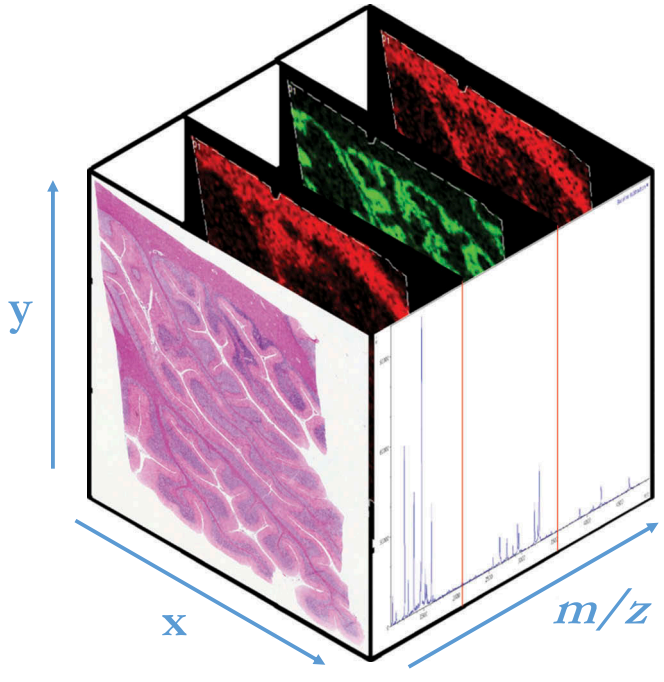

Figure 1. MALDI-MSI data cube. The x and y axes represent the spatial coordinates of the 2D digitalized tissue image (a human cerebellum tissue section is shown as an example); the z-axis represents the *mass-to-charge* (*m/z*) values in the acquired spectra. For each *m/z* value in the spectrum, a 2D molecular image is computed by coloring the pixels according to the relative abundance (intensity of that *m/z* value) of the selected compound across the tissue section. Full color available online.

Downloaded by [RMIT University Library] at 07:39 25 June 2016

3. Data elaboration

After preprocessing, which has discarded most of the technical and analytical variability within the data, the spectral dataset (in the form of the data cube) is submitted to the statistical analysis.

In this section, *machine learning* approaches are proposed to solve clinical problems, arising from the needs in the daily clinical practice, and examples of clinical applications are reported, to make the reader aware of the potentiality of the MALDI-MSI technology in the clinic.

Machine learning is the branch of computer science comprising a series of algorithms aimed at learning features from data [38] and subsequently returning the results of the inquiry performed by exploiting patterns or regularities [39] in the data [40–42]. This approach is widely employed in several fields requiring predictions from provided data, e.g. finance, computer vision, marketing, recommender systems, sentiment analysis, and search engines [43]. In biotechnology, *machine learning* has been recently implemented in numerous applications, including genetics and genomics [14,44] and proteomics [27,45], primarily aiming at finding patterns in the data for regression, classification, and clustering purposes.

Machine learning entails both supervised and unsupervised learning, according to the input data being labeled or unlabeled. The former can be addressed as the classification problem, in which a classifier is trained on labeled data in order to make predictions on unlabeled data. The latter can highlight patterns and hidden information present within the data through (mainly) clustering operations.

In the following sections, the three most common processes for solving clinical problems are described in detail: clustering, feature selection, and classification. For each, the basic statistic concepts are provided, in order to make the user aware of the operations that can be performed onto the data. Then, some examples of applications in the clinical practice are listed, to highlight the power of MALDI-MSI in aiding the clinical routine. Finally, a tutorial to achieve the proposed aims via Orange Canvas and Weka is explained.

3.1. Clustering: concepts and tools

Clustering analysis is a powerful data mining tool which does not require any previous knowledge about the data and it exploits its intrinsic properties, possibly revealing some patterns or substructures within the data [46].

The most commonly employed clustering algorithm is the hierarchical clustering (HC), which groups observations according to the similarity among each other and builds a dendrogram (Figure 2) displaying how the grouping has been performed (similar observations are placed under the same node). In MSI, a HC analysis produces, along with the dendrogram, a segmentation map (Figure 3), according to which each pixel is colored based

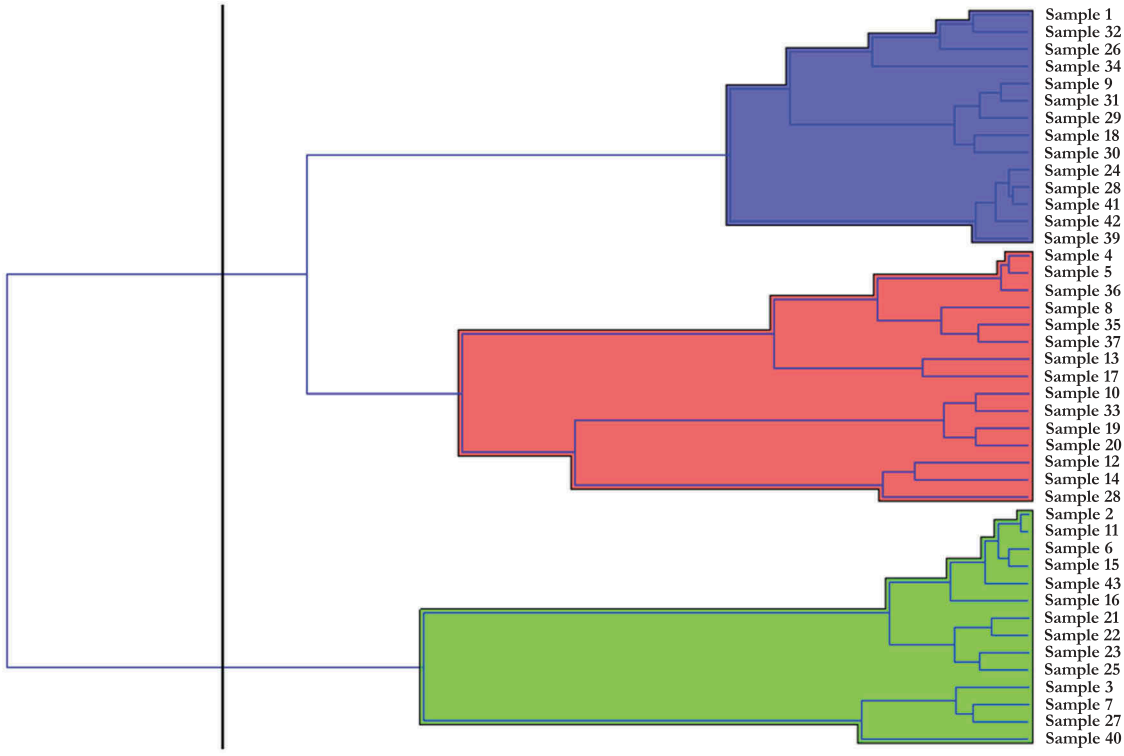

Figure 2. The figure displays an example dendrogram, in which average profiles of samples are clustered together based upon similarity only. The software (Orange Canvas in this instance) calculates the distance between samples and places similar profiles under the same node. A defined number of clusters can be set according to the distance threshold, and the software highlights the different clusters with colors. When individual spectra are used instead of the average proteomic profile, pixels corresponding to spectra under the same node are colored in the same way, resembling the color of the clusters in the dendrogram. The so-called segmentation map is therefore generated (Figure 3). However, Orange Canvas does not include a utility to generate segmentation maps, since it works on data matrices obtained after spectral elaboration through other software and does not work with spectral files directly, which retain the spatial coordinates needed to generate segmentation maps. Full color available online.

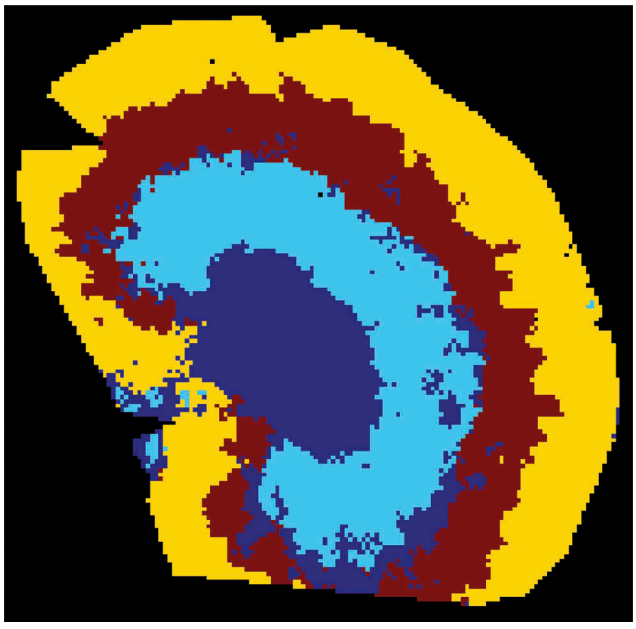

Figure 3. The figure displays an example segmentation map, obtained with SCiLS Lab 2014, onto a section of rat kidney, by coloring pixels corresponding to spectra under the same nodes with the same color. This approach has been able to correctly identify tissue sub-areas, perfectly overlapping with histo-morphological structures, without any prior knowledge. Full color available online.

upon which node the corresponding spectrum is under, thus coloring pixels referred to spectra under the same node with the same color.

3.1.1. Clinical applications

3.1.1.1. Spatially aware segmentation of neuroendocrine tumor tissue sections. In the majority of instances, however, the cluster analysis does not take into account the spatial relationship between spectra, addressing at only the signals along with their intensities. Therefore, each spectrum is considered as a completely independent observation. A spatially aware segmentation algorithm, coupled with an edge-preserving image denoising, has been developed and applied to invasive neuroendocrine tumor samples [47]. This approach addresses the problem of noisy segmentation images strictly related to noise present in the spectra, which leads to segmentation images with nondefined edges. A classical image denoising (median or convolution filter) would smooth the edges causing loss of small features in the tissue slice, which is not ideal when there are complex structures or small groups of cells to be detected. The algorithm makes use of a modified (Grasmair) total-variation-minimizing Chambolle algorithm, which minimizes the sum of absolute differences between neighbor spectra and adjusts the level of denoising to the local noise level [47]. In this way, the amount of information preserved is maximized, without discarding small features through image smoothing and better highlighting areas of interest. The edges are therefore more defined and this allows for a better delimitation of areas of clinical interest, such as tumor margins or small groups of cells: in the first instance, a clear depicting of the tumor area could be achieved, in order to surgically remove the tumor with a high degree of certainty without leaving cells in place that can possibly cause recurrence; in the latter case, a high-resolution MSI analysis could identify a very specific subpopulation of cells in order to extract information for a definition of a molecular signature for that type of cells.

3.1.1.2. Discovery of subareas of gastric cancer and human cerebellum. MALDI-MSI can highlight tissue subareas that are not always correlated with histological evidence (e.g. gastric cancer [48]) and could be used by the pathologist to better define the specimens. Moreover, tissue substructures (e.g. human cerebellum [49]) can be depicted more in detail, thanks to the possibility to overlap molecular images obtained from a MSI analysis with the same (or consecutive) tissue section after a histological staining [50]. In fact, it has been proved that HC analysis is able to highlight a possible tumor area within a tissue section, as confirmed by the pathologist after the reevaluation of the sample after staining [48]. Therefore, despite the fact that clustering is defined as an unsupervised approach, this procedure can be addressed as partially supervised, since the correct expansion of the dendrogram (number of clusters) is imposed by a histological observation. One of the biggest advantages of this approach is the potentiality to discover a tissue area of clinical interest (a tumor mass in this instance) without any previous knowledge about its presence, its position, or its molecular signature. Moreover, the fact that MALDI has the ability to ionize and detect a wide range of biomolecules further strengthens the power of this analysis.

3.1.1.3. Sarcoma intratumor heterogeneity. It is known that tumor cells undergo branching evolution within the same mass, due to genomic instability and stimuli from the surrounding microenvironment. Therefore, an individual tumor is never composed by a single type of cell, but rather by a multitude of subpopulations. This has a deep impact onto the clinical outcome of a treatment, since a drug can be effective on some subpopulations of cells but not on others, making the patient highly prone to recurrence or metastasis [51,52]. MALDI-MSI is able to detect specific protein signatures of the different subpopulations of cells that might not lead to morphological alterations. When a HC analysis is performed on the data, the dendrogram can be further explored and expanded, beyond depicting the tumor area itself, in order to reveal subareas within the same tumor region, that are histologically homogeneous [52]. Therefore, intratumor heterogeneity can be investigated, to the point where molecular signatures for each subpopulation can be generated, potentially even from tumor stem cells. As already mentioned, a MALDI-MSI dataset is highly complex, with each spectrum having a high number of features, which can make algorithms slower when computing the HC, since the calculation of the distance is more complicated when more features are employed. It has been proposed to perform the HC analysis after a step of data reduction, which operates a combination of features while preserving the information within the data [52]. *Principal Component Analysis* (PCA) generates new variables (principal components – PCs) from the linear

Downloaded by [RMIT University Library] at 07:39 25 June 2016

Downloaded by [RMIT University Library] at 07:39 25 June 2016

combination of features [53] and it is the most employed data reduction algorithm in MALDI-MSI data analysis. HC analysis performed after PCA has been able to highlight the presence of four different types of sarcoma, correlated with their clinical outcome: high-grade myxofibrosarcoma, low-grade myxofibrosarcoma, high-grade myxoid liposarcoma, and low-grade myxoid liposarcomas. High-grade myxofibrosarcoma and high-grade myxoid liposarcoma result in having a strong overlap in PCs 1 and 2, which retain the highest amount of information in the data; their separation happens in the PC 3, indicating the high degree of similarity between the two forms and the ability of MALDI-MSI in detecting small features able to discriminate between tumor species [52].

3.1.1.4. Myxofibrosarcoma tumor subpopulations. The a priori identification of tumor subpopulations is a tedious task, since, as previously mentioned, it requires the prior knowledge about the number of groups to expect. Additionally, the overlap with histology rarely produces any results, because of the lack of morphological features that characterize the subclones in the tumor [54]. Aiming at solving this issue, a multivariate statistical approach has been proposed [55,56], exploiting the multivariate nature of MSI data (many peaks associated with many compounds detected at the same time). Since each algorithm operates in its own way and can give slightly different results, the combination of a few of them has rather been implemented. The included algorithms are the following: PCA, Maximum Autocorrelation Factorization, Fuzzy C-Means, Probabilistic Latent Semantic Analysis, and Non-Negative Matrix Factorization. Since almost all of them require a value of *k* (number of populations to be expected), the algorithm has been iteratively run by ranging *k* from 2 to 10. All of these approaches imply the generation of new variables (called components) which, in turn, ensures data reduction. Consensus components were selected as the highest correlating components in these analyses in terms of score. In the end, each pixel is associated with the component that has the highest score at that location: in this manner, a segmentation map can be generated, highlighting areas corresponding to different tumor subpopulations [56]. After the identification of myxofibrosarcoma tumor subpopulations, without any previous knowledge about the clinical features of the tumor and histological match, a further study was conducted to determine the possible association between the presence of certain tumor subpopulations and the clinical outcome of the patient [56].

3.1.1.5. Classification via clustering of gastric cancer. HC can be useful when determining the classification of an unknown sample. The dendrogram can be expanded to the point where one cluster of spectra is generated for each class to show at which class the unknown sample belongs to. On the other hand, in order to do this, a large cohort of patients must be enrolled in the study and computational resources must be available to perform this operation [48]. Due to these drawbacks and to the fact that determining the correct number of classes/clusters is not automatable (since it depends on tissue histology), classification through HC is hardly performed. After determining the classification potentiality of the collected data, a classification model is rather built to achieve this purpose [48].

3.1.2. Tutorial

3.1.2.1. Orange Canvas. Orange Canvas can perform clustering analysis (Figure S1) with HC and k-means algorithms. Their parameters can be tuned via the 'Example Distance' operator, through which the distance function is set (e.g. Euclidean, correlation distance, and Manhattan), and via the clustering operator, through which the linkage function can be set.

3.1.2.2. Weka. Clustering can be performed in Weka, in the dedicated 'Cluster' tab (Figure S2). The clustering method can be selected among a variety of available algorithms, from DBSCAN to HC to k-Means algorithms. Moreover, additional parameters (such as distance function, link type, and number of clusters) can be edited for each algorithm, in order to properly tweak the clustering analysis (Figure S3).

3.2 Feature selection: concepts and tools

As mentioned previously, MALDI-MSI data, in the form of a data cube, is high dimensional and complex. Performing statistical elaboration directly onto this data can make algorithms less effective in terms of computational time and efficiency. The first step toward a reduction of the data is the *peak picking*, the feature extraction process through which only the information regarding the signals along with their intensities is preserved.

A further step in the data dimensionality reduction consists in the feature selection, namely the selection of only the features that are actually informative for the purpose to be achieved [57]. In the majority of the instances, in fact, the number of features (p) (namely, the peaks in the MSI dataset) exceeds the number of observations (n) (samples, namely, the patients), resulting in a situation that is highly prone to overfitting. In MSI, for example, individual spectra (corresponding to pixels in the digitalized image) from patients can be used instead of the average profile, in order to increase the number of observations. However, this can lead to further complications due to imbalances in the number of spectra per patient and to the lack of information that noisy and low-quality individual spectra can provide.

The feature selection process does not alter the original variables and discards all the noninformative features, which are redundant or invariant throughout the entire dataset: retaining these features would only lead to longer computational times, more noise in the data and overfitting issues when training classifiers [58].

Feature selection algorithms can be employed to prepare the data for either supervised (classification) or unsupervised (e.g. clustering) statistical analysis and they are differently implemented according to the analysis that follows. Mostly, especially in clinical applications, the data is adequately prepared for solving classification problems, since the main aim is to distinguish between benign and malignant samples in order to make predictions on unknown samples (i.e.

diagnosis). Thus, feature selection algorithms imply the use of classifiers to retain a small subset of features to characterize the biomarker discovery process.

Filter methods are more robust toward overfitting, but do not consider interactions between variables, since they do not make use of a classifier to evaluate the potentiality of the selected features. Wrapper methods evaluate the classification capability of the selected subset of features by addressing the performances of a classifier trained with those features, taking into account possible interactions between variables but being more prone to overfitting. Embedded methods include the feature selection process within the construction of the classifier, therefore being more computationally efficient than wrapper approaches [58].

3.2.1. Clinical applications

In this section, a few examples of the application of feature selection for clinical purposes are proposed. Despite not being directly related with MALDI-MSI, these approaches have been applied onto mass spectrometric data, providing high translatability to MALDI-MSI data in the selection of putative biomarkers of diagnostic or prognostic importance.

3.2.1.1. Ovarian cancer. Ovarian cancer is the second most common cancer of the female genital tract [59]. The presence of ovarian cancer is detected at a late clinical stage in more than 80% of patients, with a life expectancy of 5 years in the 35% of the cases [60]. The majority of the patients are cured by surgery alone, and the treatment is strictly dependent on the tumor subtype [59]. Classification algorithms to correctly predict the presence of ovarian cancer, in its early stages, are therefore needed, in order to increase the life expectancy of the affected patients. The feature selection is aimed at making classification algorithms faster and more efficient, and to provide a list of putative biomarkers to be used in routine diagnostic tests. One example of feature selection methodology is the employment of the *t*-test or *F*-test (analysis of variance), in order to identify the features (the peaks) that statistically vary (in intensity) between the classes of samples, discarding the invariant features [61]. This requires that the observations (spectra in the dataset) are completely independent (this is not true if multiple individual spectra from the same patient are in the dataset) and that features are normally distributed and homoscedastic, without the presence of significant outliers; moreover, *post-hoc* tests are needed when more than two classes are present, since multiple comparisons increase the chance of Type I error. When normality and homoscedasticity are not satisfied, a transformation of the data can be performed in order to meet the proper requirements [35] or nonparametric tests can be used instead [62]. The approach has been applied in the selection of a subset of peaks from mass spectra to be used in the classification (via Support Vector Machine (SVM), Random Forests (RF), and k-Nearest Neighbor (k-NN) classifiers) of ovarian cancer bioptic samples [61].

3.2.1.2. Leukemia blood sera. A Bayesian inductive method has been proposed in the selection of relevant peaks from mass spectra acquired from blood sera of patients affected by leukemia [63]. Despite being model-independent, this method is capable of detecting relationships between spectral features, via the employment of the concept of mutual information when determining the impact of the feature on the classification. The Bayesian network/mutual information approach leads to a selection of a small subset of features that decreases the risk of overfitting and provides a reduced list of mass values to be investigated as potential biomarkers. Moreover, the feature subset has been used for the construction of a model that accurately makes predictions on new data, adding clinical relevance to the results [63].

3.2.2. Tutorial

3.2.2.1. Orange Canvas. Feature selection can be performed in Orange Canvas only through Python scripting. The software, in fact, does not implement a graphical widget for the feature selection process, but includes a Python module (called 'selection') that can be loaded when scripting. Although this requires programming skills, the process is well explained in the software documentation [64]. The only way to select features in Orange Canvas, however, is through the VizRank widget [65], which finds the best data projections to separate data points of different classes. In order to achieve this, the best projections are established by evaluating the classification performances of a trained k-NN model. The software allows for a selection of a maximum number of features to be employed in the evaluation, so that a feature selection is performed on the data before generating the projections (Figure S4).

3.2.2.2. Weka. Weka software has a dedicated section ('Select attributes') where to perform feature selection. It is in fact possible to select the method to be used when selecting features (e.g. chi-squared evaluation and evaluation of subsets through filter or wrapper algorithms) and it is possible to set the parameters for the feature selection process (Figure S5 and Figure S6). Moreover, the feature selection method can be selected (e.g. forward, backward, bidirectional, and stepwise) with additional parameters for each.

3.3. Classification: concepts and tools

The classification problem is one of the major instances under the supervised learning, and it is aimed at assigning unknown samples to a specific class according to the information provided by their features [66]. In order to achieve this, a classifier must be trained onto labeled samples (training dataset, consisting of several observations of known class), to compute the mathematical functions that explain the relationship between the features (explanatory variables) and the class (response variable). In order to assess the classification capability of the trained classifier, a cross-validation can be performed onto the training dataset itself, by iteratively splitting it into two subsets to be used, in turn, as training and validation subsets. Furthermore, it is possible to test the classifier performances onto an external validation dataset, by evaluating the discrepancy between the predicted class and the actual class.

The classification problem is present in several applications, such as computer vision, speech recognition, and biometric identification, with different algorithms employed for the purpose of classifying unknown samples. In biology, the

Downloaded by [RMIT University Library] at 07:39 25 June 2016

Downloaded by [RMIT University Library] at 07:39 25 June 2016

classification problem represents the ability of the analytical approach to reliably discriminate between samples under different conditions (e.g. benign and pathological, stage of the disease, treatment conditions, etc.) [67–71].

In most of the instances, the diagnosis is performed via the evaluation of a histologically stained bioptic tissue section retrieved from the patient by pathologists and it is strongly dependent from their training and experience in order to detect smaller features across the tissue section. Moreover, subtle molecular changes directly correlated with morphological modifications cannot be appreciated by the human eye. Therefore, this results in many samples being filed as undetermined reports [72] or being addressed as pathological only in the late stages [67]. MALDI-MSI technology, by looking at the sample at the molecular level, can detect small molecular changes already in the early stages of the disease, even when the tissue looks morphologically healthy [73].

An example algorithm that is widely used as classifier in many applications is the SVM [74]. Models using SVMs are computed by fitting a single (or a set of) hyperplane(s), in a high-dimensional space, which maximize the minimal distance between data points belonging to different classes. SVMs can go beyond linear classification, as they can be used as non-linear classifiers, thanks to the *kernel function* which allows the switch to a transformed feature space for better fitting the separating hyperplane(s) [74]. The algorithm has been implemented in many biological applications, directly exploiting MALDI-MSI data, with successful results.

RF algorithms are ensemble decision tree methods characterized by being robust to overfitting and by providing high prediction accuracy, guaranteeing high performances even with a large input dataset [75].

3.3.1. Clinical applications

3.3.1.1. Breast cancer. Breast cancer constitutes one of the main causes of mortality among women. The presence of the human epidermal growth factor receptor 2 (HER2) has been found to be strictly correlated with the response to the treatment with trastuzumab (herceptin) and with the clinical outcome of the patient. Therefore, the reliable assessment of the presence of HER2 is of high clinical importance [68]. A SVM classifier has been able to correctly detect the presence of HER2 onto human breast cancer samples, with high values of accuracy, sensitivity, and specificity [68]. RF algorithms, on the other hand, have been successfully applied onto MALDI-MSI data for the classification of breast tumor cells in xenograft mouse models [69]. Proteomic profiles (average spectra from histologically selected regions of interest) of different subregions (necrotic tissue, tumor mass, gelatine, tumor interface, and no tissue) of the tumor have been generated and passed to the classifier for training. Since the algorithm entails an ensemble method, the process results in being more robust and reliable, since the classification outcome is the result of a vote among the classifiers built within the ensemble [69].

3.3.1.2. Metastasis identification. Metastases are defined as tumor cells detaching from the original tumor mass, invading the surrounding environment and populating an organ that is different from the organ of origin. However, in many cases, metastases cannot be associated to any primary tumor mass (cancer of unknown primary (CUP)) making the treatment more difficult, due to this lack of information [70]. The analysis of known tumor masses can provide the molecular signature of each type of tumor, allowing to train a classifier that is able to assign unknown metastases to the primary tumor mass. This approach has led to a confident determination of the type of primary tumor related to the metastases, in such a way that a more defined treatment for each patient can be established [70]. The RF and SVM classifiers have been trained onto features coming from the proteomic profiles of the different types of primary tumor, in order to predict the class of metastasis coming from unknown primary tumor masses (CUP) [70].

3.3.1.3. Disease progression: liver cirrhosis and metastatic melanoma. MALDI imaging can be useful in determining the stage of the disease [67,71], in order to operate the diagnosis at the very early stages or to modify the treatment in accordance with the clinical outcome of the disease. MALDI-MSI has been employed in predicting if a condition of liver cirrhosis is evolving toward cancer (hepatocellular carcinoma (HCC)) [67]. The malignancy of a cirrhosis is often determined when the disease is at its late stages and in some instances the cirrhotic tissue that remains after the surgical removal of the tumor can cause recurrences. This is a clear example of the potentiality of MALDI-MSI in predicting the intrinsic nature of a liver cirrhosis and preventing the evolution of the patient toward a poor prognosis. Representative spectra of cirrhosis without HCC, cirrhosis with HCC, and HCC have been collected from the analysis of specimens, and a SVM classifier has been trained to correctly predict the clinical evolution of unknown specimens. Finally, the presence of Ubi(1-74) (a truncated form of ubiquitin) has been found to be strictly correlated with the clinical outcome of the patients, providing a suitable target for immunohistochemical tests to be used in the clinical routine [67].

Protein signature of tumor recurrence has also been generated by MALDI-MSI in order to predict the stage of metastatic melanoma, through the analysis of lymph nodes [71]. Proteins (histone H4, cytochrome c, thymosin, and ubiquitin) correlated with the patient outcome have been identified [71] in order to provide a putative biological meaning and, again, to translate the findings to a diagnostic test to be employed in the routine clinical practice. The stage of melanoma is usually determined by histological evaluation of the tumor features according to the general established guidelines, but the correlation with the prognosis is often compromised by the multitude of features to evaluate [71]. MALDI-MSI provides a molecular insight on the disease, by addressing the problem in a multivariate way onto molecular bases. Proteins differently expressed between healthy lymph nodes and metastatic melanoma lymph nodes have been selected through a Significance Analysis of Microarrays (SAM) test, yielding a set of signals constituting a molecular signature of prognosis. An ensemble of four models (Genetic Algorithm, SVM, Supervised Neural Network, and Quick Classifier) has been trained onto the selected features, and tested for robustness and

performance assessment, obtaining high accuracy in classifying bioptic samples [71].

3.3.1.4. Tumor margins. Tumor recurrence is the most dangerous consequence after the treatment of a tumor, since it is often more aggressive and chemo-resistant, compromising the treatment of the patient: the main cause of recurrence is left-over tumor cells, which are indistinguishable under light microscopy after histological staining, and are therefore left in place to repopulate the tumor mass [76,77]. MALDI-MSI can highlight the presence of tumor cells at the molecular level, by directly guiding the surgery or by acting as advisor to the physician [76]. In this context, MALDI-MSI has been able to successfully detect left-over sarcoma [76] and clear cell renal cell carcinoma [77] cells after surgical treatment. The selection of features (peaks in the mass spectra), which act as a signature of malignancy, has been performed by using the SAM and permutation *t*-test for paired data and by picking only the features with a false discovery rate less than 0.01. The discriminatory capability of the selected features has been assessed by training a classifier (SVM) and evaluating its performances. This further proves the value of MALDI-MSI in going beyond the morphological evaluation after staining and in providing clinical translatability when identifying discriminatory compounds to be employed in a routine test (such as immunohistochemistry) [76,77].

3.3.2. Tutorial

3.3.2.1. Orange Canvas. Orange Canvas provides a variety of tools for classification problem solving (Figure S7). The input data can be split into training and test set by setting a filter condition (e.g. a threshold in a feature value or a sample name) and multiple classifiers can be trained and tested at the same time, both via a k-fold cross-validation onto the training data and via its application onto an external test set. The classification performances can be evaluated through many parameters (such as sensitivity, specificity, accuracy, predictive values, Receiver Operating Characteristic analysis, and so on), providing a detailed report onto the classifier's behavior.

3.3.2.2. Weka. Weka can be used to solve classification problems through its implementation of a classification system, available in the 'Classify' tab (Figure S8). Weka includes a great variety of classifiers, such as Bayes classifiers, SVMs, linear regression, PLS, logistic regression, and RFs. In addition, it is possible to tune the classifier parameters in order to maximize its classification capability: for example, for the SVM, it is possible to set cost, kernel function, degree of the polynomial, epsilon, gamma, seed (for pseudo-randomization), weights, and nu (Figure S8). Finally, Weka allows the selection of the performance assessment method: test onto the entire dataset (using the training dataset as a test set), test onto an external validation set, k-fold cross-validation onto the entire dataset, and train/test split. It then returns the detailed report for the classification performances, in terms of sensitivity, specificity, and accuracy, along with the confusion matrix and other data. Classification can be performed after a feature selection process, by manually preserving only the features that are listed as significant in the feature selection output within Weka (Figure S5).

4. Expert commentary

Data dimensionality still represents a big issue in terms of computational efficiency and storage of the acquired data, and statistical methods of information-preserving data reduction constitute a key point in the data mining and elaboration phase. The main aim is to provide a reliable and informative output in reasonable time, without discarding any important features from the data.

Traditional inference tasks, such as clustering, feature selection, or classification, attempt to find patterns in a dataset characterized by a collection of independent instances of a single table. Numerous algorithms have been designed to work on such a standard approach, where instances can be easily represented as fixed-length vectors of attribute values. Unfortunately, many studies still do not consider that real problems are best described by structured data where instances of multiple types are related to each other in complex ways. For this reason, datasets to be analyzed may be described by relational databases or semi-structured representations such as XML. In this case, features of one entity are often correlated with features of related entities. It may happen that, just as some features are not helpful for mining datasets, some relations might provide information for clustering or classification algorithms. For instance, when it comes to analyzing differentially expressed MS peaks in a case–control classification problem, comparisons are generally performed between protein/peptide profiles of different groups or between statistics summarizing the peak properties of a group. In such a situation, the incorporation of relational information can give powerful (case–control) discriminatory capability. This has been proved useful in many fields [28] [78–80] and represents a promising approach also in relation of both Multidimensional Protein Identification Technology data structure and MS improvement, in instruments and methods, such as targeted proteomics or data-independent analysis.

5. Five-year view

MALDI-MSI is an analytical technique that is characterized by being versatile and highly translatable to the daily clinical practice. The high sensitivity of the technique, coupled with high specificity and high spatial resolution, makes it a valuable resource in aiding diagnoses, by providing a molecular insight of the specimen. The possibility to integrate data derived from a MALDI-MSI analysis with the most common clinical techniques (such as immunohistochemistry and histology) increases interoperability and the reliability of MSI data in being used in the daily clinical routine. However, the potentiality of the employment of this technology in solving clinical problems and further supporting the daily clinical routine diagnosis is strongly dependent on the data storage and elaboration. At the present time, two different aspects are critical in its application, beyond technological and methodological optimization: hardware and software. Faster and new design CPU and

Downloaded by [RMIT University Library] at 07:39 25 June 2016

storage devices to speed up data elaboration and to keep the huge number of mass spectra will be really welcome. On the other side, new and better performing algorithms are needed in order to reduce the amount of time to be dedicated to the data elaboration processes, to decrease the manual intervention of the personnel, and to make robust, automatic and easy-to-use (also by not experts) software.

Key issues

- Matrix-Assisted Laser Desorption/Ionization – Mass Spectrometry Imaging (MALDI-MSI) is able to provide a molecular insight of the samples, detecting the presence of a great variety of analytes directly *on-tissue* and showing their spatial distribution across the tissue section.
- A typical MALDI-MSI dataset is composed of mass spectra corresponding to pixels of the digitalized tissue slice and it is structured as a data cube, in which every *mass-to-charge* ratio (*m/z*) value is associated with a molecular image showing the localization of that specific analyte *on-tissue*. The dimensionality of the data strictly depends on the mass resolution and on the spatial distribution (i.e. number of pixels) of the spectral acquisition.
- The preprocessing phase ensures that all the spectra of the dataset are brought to the same scale, allowing fair comparisons between spectra/pixels within the same tissue section and among different analyses, by discarding all the fluctuations associated with instrument performances and sample heterogeneity.
- *Machine learning* comprises a series of algorithms aimed at learning features from data and subsequently returning the results by exploiting patterns or regularities within the data. This approach is widely employed in several fields, for clustering (unsupervised) and classification (supervised) purposes. While the former do not require any prior knowledge about the label of the data and return hidden patterns within the data, the latter exploit the known input data in order to make predictions onto new unlabeled data.
- Clustering analysis is a powerful data mining tool, that exploits the intrinsic properties of the data to reveal some patterns or substructures within it. One of the biggest advantages of this unsupervised approach is that it does not require any previous knowledge about the data, but it can highlight sub-groups of observations (i.e. mass spectra) that can become of clinical interest. In mass spectrometry imaging, clustering analysis is associated with segmentation maps, coloring pixels referred to spectra under the same node with the same color and thus depicting subareas of possible high clinical importance.
- Feature selection discards all the non-informative features, that are redundant or invariant throughout the entire dataset, fully preserving the original variables without any mathematical operations on the values: by doing this, shorter computational times are achieved, along with a lower tendency to overfitting when training classifiers. Feature selection algorithms that are employed for solving classification problems (e.g. diagnosis) make use of classifiers to retain a small subset of features to characterize the biomarker discovery process: the list of preserved features, in fact, may constitute a molecular signature of malignancy, to be exploited by clinical diagnostic tests.
- The classification problem, one of the major instances under the supervised learning, represents the ability of the analytical approach to discriminate between samples under different conditions (e.g. benign and pathological, stage of the disease, treatment conditions, etc…). Several classifiers (such as Support Vector Machine – SVM and Random Forests – RF) have been trained on MALDI-MSI data for the accurate classification of unknown samples coming from the clinical routine, in order to exploit the potentiality of the technology to look at the molecular level to reliably aid the diagnostic process.
- MALDI-MSI has proven its capability in assisting the daily clinical routine by providing a molecular view of the specimens, revealing subtle molecular changes that may not be directly correlated with morphological modifications that can be evaluated by pathologists, especially in the early stages of the disease. Therefore, this will result in less samples being filed as undetermined reports or being addressed as pathological only in the late stages.

Declaration of interest

The authors declare the following funding information MIUR: FIRB 2007 (RBRN07BMCT 11); FAR 2010-2014; iMODE-CKD (FP7-PEOPLE-2013-ITN-608332); in part the COST Action (BM1104) Mass Spectrometry Imaging: New Tools for Healthcare Research; Fondazione Gigi & Pupa Ferrari Onlus. The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

ORCID

Manuel Galli http://orcid.org/0000-0003-4862-9599
Italo Zoppis http://orcid.org/0000-0001-7312-7123
Fulvio Magni http://orcid.org/0000-0002-8663-0374
Giancarlo Mauri http://orcid.org/0000-0003-3520-4022

References

Papers of special note have been highlighted as:
• of interest
•• of considerable interest

1. Caprioli RM, Farmer TB, Gile J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. Anal Chem. 1997;69(23):4751–4760.
2. Lalowski M, Magni F, Mainini V, et al. Imaging mass spectrometry: a new tool for kidney disease investigations. Nephrol Dial Transpl. 2013;28(7):1648–1656.
3. Aichler M, Walch A. MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. Lab Invest. 2015;95(4):422–431.
4. Gorzolka K, Walch A. MALDI mass spectrometry imaging of formalin-fixed paraffin-embedded tissues in clinical research. Histol Histopathol. 2014;29(11):1365–1376.
5. Kriegsmann J, Kriegsmann M, Casadonte R. MALDI TOF imaging mass spectrometry in clinical pathology: A valuable tool for cancer diagnostics (review). Int J Oncol. 2015;46(3):893–906.
6. Cole LM, Clench MR. Mass spectrometry imaging for the proteomic study of clinical tissue. Proteomics Clin Appl. 2015;9(3–4):335–341.

Downloaded by [RMIT University Library] at 07:39 25 June 2016

Downloaded by [RMIT University Library] at 07:39 25 June 2016

7. Trim PJ, Snel MF. Small molecule MALDI MS imaging: current technologies and future challenges. Methods. 2016. pii:S1046–2023(16)30011-1. [Epub ahead of print].
8. Trim PJ, Francese S, Clench MR. Imaging mass spectrometry for the assessment of drugs and metabolites in tissue. Bioanalysis. 2009;1:309–319.
9. Trede D, Kobarg JH, Oetjen J, et al. On the importance of mathematical methods for analysis of MALDI-imaging mass spectrometry data. J Integr Bioinform. 2012;9(1):189.
10. Mierswa I, Wurst M, Klinkenberg R, et al. YALE: rapid prototyping for complex data mining tasks. Proc ACM SIGKDD Int Conf Knowl Discov Data Min. 2006;2006:935–940.
11. Verbeeck N, Yang J, De Moor B, et al. Automated anatomical interpretation of ion distributions in tissue: linking imaging mass spectrometry to curated atlases. Anal Chem. 2014;86(18):8974–8982.
12. Römpp A, Spengler B. Mass spectrometry imaging with high resolution in mass and space. Histochem Cell Biol. 2013;139(6):759–783.
13. Zoppis I, Merico D, Antoniotti M, et al. Discovering relations among GO-annotated clusters by graph Kernel Methods. In: Măndoiu I, Zelikovsky A, editors. Bioinformatics research and applications (Proceedings of third International Symposium, ISBRA 2007; 2007 May 7–10; Atlanta, GA). Berlin: Springer; 2007. p. 158–169.
14. Cava C, Zoppis I, Gariboldi M, et al. Combined analysis of chromosomal instabilities and gene expression for colon cancer progression inference. J Clin Bioinf. 2014;4(1):2.
15. Antoniotti M, Carreras M, Antonella F, et al. An application of kernel methods to gene cluster temporal meta-analysis. Comput Oper Res. 2010;37(8):1361–1368.
16. The MathWorks Inc. MATLAB version R2015b [Internet]. Natick (MA): The MathWorks Inc; 2015. Available from: http://www.mathworks.com/products/connections/?refresh=true
17. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna: R Core Team; 2016.
18. Ketterlinus R, Hsieh SY, Teng SH, et al. Fishing for biomarkers: analyzing mass spectrometry data with the new ClinProTools software. BioTechniques. 2005;38(S6):37–40.
19. [cited 2016 Jun 16]. Available from: http://scils.de/.
20. Schramm T, Hester A, Klinkert I, et al. ImzML - A common data format for the flexible exchange and processing of mass spectrometry imaging data. J Proteomics. 2012;75(16):5106–5110.
21. Demšar J, Curk T, Erjavec A, et al. Orange: data mining toolbox in python. J Mach Learn Res. 2013;14(1):2349–2353.
22. [cited 2016 Jun 16]. Available from: http://orange.biolab.si/
23. Demšar J, Zupan B, Leban G, et al. Orange: from experimental machine learning to interactive data mining. Berlin: Springer; 2004.
24. Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. SIGKDD Explor. 2009;11(1):10–18.
25. Available at: http://www.cs.waikato.ac.nz/ml/weka/ [Last accessed 16 June 2016]
26. Jupp S, Eales J, Fischer S, et al. "Combining rapidminer operators with bioinformatics services - a powerful combination," In *RapidMiner Community Meeting and Conference*, 2011.
27. Zoppis I, Antoniotti M, Mauri G, et al. Mutual information optimization for mass spectra data alignment. IEEE/ACM Trans Comput Biol Bioinforma. 2012;9(3):934–939.
28. Zoppis I, Borsani M, Gianazza E, et al. Analysis of correlation structures in renal cell carcinoma patient data. BIOINFORMATICS 2012 - Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms; 2012 Feb 1–4; Vilamoura. p. 251–256.
29. Goodwin RJA. Sample preparation for mass spectrometry imaging: small mistakes can lead to big consequences. J Proteomics. 2012;75(16):4893–4911.
30. Schwartz SA, Reyzer ML, Caprioli RM. Direct tissue analysis using matrix-assisted laser desorption/ionization mass spectrometry: practical aspects of sample preparation. J Mass Spectrom. 2003;38(7):699–708.
31. Norris JL, Cornett DS, Mobley JA, et al. Processing MALDI mass spectra to improve mass spectral direct tissue analysis. Int J Mass Spectrom. 2007;260(2–3):212–221.
32. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. Anal Chem. 1964;36(8):1627–1639.
33. Gan F, Ruan G, Mo J. Baseline correction by improved iterative polynomial fitting with automatic threshold. Chemom Intell Lab Syst. 2006;82(1–2) SPEC. ISS:59–65.
34. Deininger SO, Cornett DS, Paape R, et al. Normalization in MALDI-TOF imaging datasets of proteins: practical considerations. Anal Bioanal Chem. 2011;401(1):167–181.
• **It shows how data preprocessing can affect the results by generating possible artifacts.**
35. Wolski WEWE, Lalowski M, Martus P, et al. Transformation and other factors of the peptide mass spectrometry pairwise peak-list comparison process. BMC Bioinfo. 2005;6(1):285.
36. Yang C, He Z, Yu W. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. BMC Bioinfo. 2009;10:4.
37. Alexandrov T, Meding S, Trede D, et al. Super-resolution segmentation of imaging mass spectrometry data: solving the issue of low lateral resolution. J Proteomics. 2011;75(1):237–245.
38. Abu-Mostafa YS, Magdon-Ismail M, Lin H-T. Learning from data, no. 2. AMLBook; 2012.
39. Theodoridis S, Koutroumbas K. Pattern recognition. 3rd ed. Vol. 11. Cambridge (MA): Academic Press; 2006.
40. Bishop CM. Pattern recognition and machine learning. Vol. 4. no. 4. New York: Springer-Verlag; 2006.
41. Mitchell TM. Machine learning, International edition. 1st ed. New York City: McGraw-Hill Education; 1997.
42. Pang-Ning T, Steinbach M, Kumar V. Introduction to data mining. 1 ed. London: Pearson; 2005.
43. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Elements. 2009;1:337–387.
44. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321–332.
45. Kelchtermans P, Bittremieux W, De grave K, et al. Machine learning applications in proteomics research: how the past can boost the future. Proteomics. 2014;14(4–5):353–366.
46. Jain AK, Dubes RC. Algorithms for clustering data. Prentice Hall. 1988;355:320.
47. Alexandrov T, Becker M, Deininger SO, et al. Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. J Proteome Res. 2010;9(12):6535–6546.
• **It describes how a full preservation of the spatial information better highlights areas of interest that are used to enhance the image of the sample.**
48. Deininger SO, Ebert MP, Fütterer A, et al. MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. J Proteome Res. 2008;7(12):5230–5236.
49. De Sio G, Smith AJ, Galli M, et al. A MALDI-mass spectrometry imaging method applicable to different formalin-fixed paraffin-embedded human tissues. Mol Biosyst. 2015;11(6):1507–1514.
•• **The versatility of MALDI-MSI in being applicable to different types of tissue and the intrinsic potentiality of the acquired data are well described.**
50. Schwamborn K, Krieg RC, Reska M, et al. Identifying prostate carcinoma by MALDI-imaging. Int J Mol Med. 2007;20(2):155–159.
51. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer. 2012;12(5):323–334.
52. Willems SM, Van Remoortere A, Van Zeijl R, et al. Imaging mass spectrometry of myxoid sarcomas identifies proteins and lipids specific to tumour type and grade, and reveals biochemical intra-tumour heterogeneity. J Pathol. 2010;222(4):400–409.
53. Yao I, Sugiura Y, Matsumoto M, et al. In situ proteomics with imaging mass spectrometry and principal component analysis in the scrapper-knockout mouse brain. Proteomics. 2008;8(18):3692–3701.
54. Wu JM, Halushka MK, Argani P. Intratumoral heterogeneity of HER-2 gene amplification and protein overexpression in breast cancer. Hum Pathol. 2010;41(6):914–917.

55. Jones EA, van Remoortere A, van Zeijl RJM, et al. Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma. PLoS One. 2011;6(9):e24913.
56. Balluff B, Frese CK, Maier SK, et al. De novo discovery of phenotypic intratumour heterogeneity using imaging mass spectrometry. J Pathol. 2015;235(1):3–13.
•• **A real example of the diagnostic power of MALDI-MSI data, where the intra-tumour heterogeneity can have a deep impact on the outcome of a patient after treatment.**
57. Guyon I, Gunn S, Nikravesh M, et al. Feature extraction: Foundations and applications. 1st ed. Vol. 207. Berlin: Springer-Verlag; 2006.
58. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–2517.
59. Ramalingam P. Morphologic, immunophenotypic, and molecular features of epithelial ovarian cancer. Oncology (Williston Park). 2016;30(2):1–15.
60. Barakat RR, Markman M, Randall M. Principles and practice of gynecologic oncology. Vol. 16. no. 3. Philadelphia (PA): Lippincott Williams & Wilkins; 2009.
61. Datta S, Depadilla LM. Feature selection and machine learning with mass spectrometry data for distinguishing cancer and non-cancer samples. Stat Methodol. 2006;3(1):79–92.
62. Yu JS, Ongarello S, Fiedler R, et al. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. Bioinformatics. 2005;21(10):2200–2209.
63. Kuschner KW, Malyarenko DI, Cooke WE, et al. A bayesian network approach to feature selection in mass spectrometry data. BMC Bioinfo. 2010;11:177.
64. [cited 2016 Jun 16]. Available from: http://docs.orange.biolab.si/2/reference/rst/Orange.feature.selection.html
65. Leban G, Bratko I, Petrovic U, et al. VizRank: finding informative data projections in functional genomics by machine learning. Bioinformatics. 2005;21(3):413–414.
66. Duda RO, Hart PE, Stork DG. Pattern classification. New York, NY: John Wiley, Section; 2000. p. 654.
67. Laouirem S, Le Faouder J, Alexandrov T, et al. Progression from cirrhosis to cancer is associated with early ubiquitin post-translational modifications: identification of new biomarkers of cirrhosis at risk of malignancy. J Pathol. 2014;234(4):452–463.
68. Rauser S, Marquardt C, Balluff B, et al. Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. J Proteome Res. 2010;9(4):1854–1863.
69. Hanselmann M, Köthe U, Kirchner M, et al. Toward digital staining using imaging mass spectrometry and random forests. J Proteome Res. 2009;8(7):3558–3567.
70. Meding S, Nitsche U, Balluff B, et al. Tumor classification of six common cancer types based on proteomic profiling by MALDI imaging. J Proteome Res. 2012;11(3):1996–2003.
71. Hardesty WM, Kelley MC, Mi D, et al. Protein signatures for survival and recurrence in metastatic melanoma. J Proteomics. 2011;74(7):1002–1014.
72. Pagni F, Prada M, Goffredo P, et al. 'Indeterminate for malignancy' (Tir3/Thy3 in the Italian and British systems for classification) thyroid fine needle aspiration (FNA) cytology reporting: morphological criteria and clinical impact. Cytopathology. 2014;25(3):170–176.
73. Pagni F, Mainini V, Garancini M, et al. Proteomics for the diagnosis of thyroid lesions: preliminary report. Cytopathology. 2015;26(5):318–324.
74. Cristianini N and Shawe-Taylor J. *An introduction to Support Vector Machines*, vol. 47, no. 2. Cambridge, UK: Cambridge University Press, 2000.
75. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
76. Caldwell RL, Gonzalez A, Oppenheimer SR, et al. Molecular assessment of the tumor protein microenvironment using imaging mass spectrometry. Cancer Genomics Proteomics. 2006;3:279–288.
77. Oppenheimer SR, Mi D, Sanders ME, et al. Molecular analysis of tumor margins by MALDI mass spectrometry in renal carcinoma. J Proteome Res. 2010;9(5):2182–2190.
• **Another example of clinical application where MALDI-MSI can be used to aid surgery in order to avoid recurrence.**
78. Kolaczyk E. Statistical analysis of network data. New York: Springer; 2009.
79. Pekalska E, Duin RPW. The dissimilarity representation for pattern recognition : foundations and applications. Ser Mach Percept Artif Intell. 2005;64:xxvi, 607.
80. Long B, Zhang Z, Yu PS. Relational data clustering: models, algorithms, and applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Virginia Beach (VA): Chapman and Hall/CRC; 2010.

Downloaded by [RMIT University Library] at 07:39 25 June 2016