The University of Milano Bicocca

THE PhD PROGRAM DIMET

The PhD Program in Translational and Molecular Medicine (DIMET) is an inter-departmental project between the School of Medicine and the School of Science, organized by the University of Milano-Bicocca.

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

DIMET - Dr. Manuel GALLI - A.A. 2016-17

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

PhD

PROGRAM IN TRANSLATIONAL
AND MOLECULAR MEDICINE

DIMET

UNIVERSITY OF MILANO-BICOCCA
SCHOOL OF MEDICINE AND SCHOOL OF SCIENCE

An easy-to-use software program for the ensemble pixel-by-pixel classification of MALDI-MSI datasets

Coordinator: Prof. Andrea Biondi
Tutor: Prof. Fulvio Magni
Co-Tutor: Dr. Italo Zoppis

Dr. Manuel GALLI

Matr. No. 717890

XXX CYCLE
ACADEMIC YEAR
2016-2017

SCUOLA DI DOTTORATO

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of

MEDICINE AND SURGERY

PhD Program

TRANSLATIONAL AND MOLECULAR MEDICINE (DIMET)

Cycle **XXX**

AN EASY-TO-USE SOFTWARE PROGRAM FOR THE ENSEMBLE PIXEL-BY-PIXEL CLASSIFICATION OF MALDI-MSI DATASETS

Surname: **GALLI** Name: **MANUEL**

Registration Number: **717890**

Tutor: **PROF. FULVIO MAGNI**

Co-Tutor: **DR. ITALO ZOPPIS**

Coordinator: **PROF. ANDREA BIONDI**

ACADEMIC YEAR 2016/2017

Statistics only explains the variance within the data…

It does not explain emotions,

and it never will…

Table of Contents

Chapter 1

Introduction

1.1 Background and state of the art

1.1.1 Mass Spectrometry Imaging (MSI)

The mass spectrometer

Mass spectrometry is an analytical technique which involves the use of an advanced instrumentation, the mass spectrometer, in order to obtain information regarding the molecular composition of a specific sample, such as the presence and abundance of a great variety of molecules and the structure of molecules of interest [1].

In order to do so, the mass spectrometer generates mass spectra, via the ionization of the molecules present in the sample and the determination of their *mass-to-charge ratio* (*m/z*) by the exploitation of electromagnetic fields [1].

The instrument is composed of different compartments: a sample inlet system, an ion source, a mass analyzer and a detector. Each one of them is an independent unit: therefore, a variety of mass spectrometers can be assembled, depending on the application.

The ion source is the chamber in which gas-phase ions are generated from the sample. There are different types of ion sources and each one of them uses a different technique to generate ions. The most important influencing aspects of the ionization are the internal energy transferred during the process and the physico-chemical properties of the analyte. Some ionization techniques are very energetic and cause the fragmentation of the analyte; other techniques are softer and only produce ions of molecular species. Among the most employed ion sources are *ElectroSpray Ionization* (ESI) [2], *Matrix-Assisted Laser Desorption/Ionization* (MALDI) [3], *Secondary Ion Mass Spectrometry* (SIMS) [4] and *Desorption ElectroSpray Ionization* (DESI) [5].

The mass analyzer is the compartment of the mass spectrometer that separates gas-phase ions (obtained from the sample in the ion source) according to their *mass-to-charge ratio* (*m/z*). Each analyzer is characterized by different parameters: the mass range of measurement, the speed of analysis, the transmission rate (ratio of the number of ions reaching the detector to the number of ions entering the mass analyzer), the mass accuracy and the resolving power (ability of the analyzer of separating two ions with a small difference in their *m/z* ratios, generating two distinct peaks in the spectrum).

Among the most employed mass analyzers are the *quadrupole* [6], the 2D and 3D *ion trap* [7], the *Orbitrap* [8], the *FT-ICR* (*Fourier Transform - Ion Cyclotron Resonance*) [9] and the *TOF* (*Time Of Flight*) analyzer [10].

Finally, after being generated in the source and traveling across the analyzer, the ions hit the detector, which allows the generation of an electric current, directly proportional to the amount of ions hitting the detector plate. The current is then amplified and converted to digital by the digitizer, which yields the mass spectrum, recording the presence and relative abundance of ions [1].

The MALDI-TOF instrumentation

One of the most commonly used mass spectrometers, especially for imaging approaches, is the Matrix-Assisted Laser Desorption/Ionization - Time Of Flight (MALDI-TOF) mass spectrometer. This is because it provides a good balance between the ease of the sample preparation and the versatility of the instrumentation in performing the analysis. The MALDI source, in fact, is able to ionize a wide variety of molecules (proteins, peptides, lipids, oligonucleotides, polymers and inorganic compounds), and this makes it suitable for the analysis of different compounds, from chemicals to biological molecules. The TOF analyzer guarantees high speed of analysis, high versatility in terms of mass range and high transmission efficiency.

In order to achieve the ionization through MALDI, the analytes must be firstly dissolved in a solution containing a particular chemical compound, the **matrix**, which, when the solvent evaporates, co-crystallizes with the analyte molecules, extracting them from the sample, resulting in a solid deposit on the target. The desorption process takes place in the spectrometer source under high vacuum. A laser (emitting light at a certain wavelength) hits the sample with an intense pulse, causing the detachment of some portions of this solid deposit: this process is triggered by the fact that the matrix molecules absorb energy from the laser and transfer this energy to the analyte molecules, which are now in the gas phase, ionized.

The TOF (Time Of Flight) mass analyzer separates ions according to their velocity, namely the time that ions take to cover the length of the flight tube, which is free from electromagnetic fields, under high vacuum. Ions generated by the source are previously accelerated by a potential difference applied between an electrode and the extraction grid. Every ion has got the same electric potential energy, which is entirely converted into kinetic energy: therefore the velocity of an ion is inversely proportional to its *mass-to-charge ratio* (*m/z*).

The mass spectrum

The analyte molecules present in the sample are ionized in the source and separated according to their *m/z* in the mass analyzer. The detector records the impact of the ions and translates this into an electric signal, which is then converted into digital by the digitize. The final output is a *m/z* x intensity graph, a mass spectrum (Figure 1), which displays the presence and the relative abundance of molecules present in the sample.

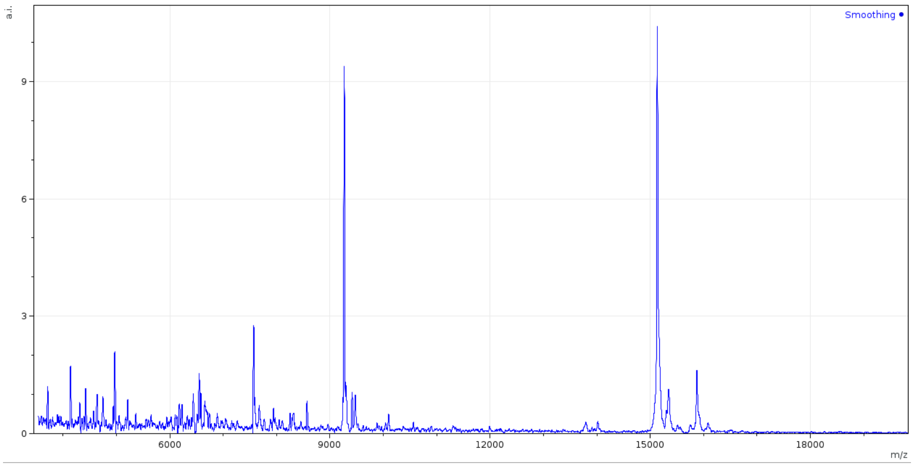


Figure 1: Example of a mass spectrum. The x-axis represents the mass-to-charge ratio (m/z) of the molecules present in the sample, while the y-axis represents the relative abundance of those molecules.

The MALDI-MSI approach

The MALDI-TOF mass spectrometric approach allows to obtain a molecular profile, representative of the sample, which can be employed for biomarker research aimed at diagnosis, patient classification or disease understanding. This approach, also referred to as Mass Spectrometry Profiling (MSP), is mostly suitable for the analysis of fluids (such as blood or urine) or tissue homogenates, where a representative aliquot of the sample is analyzed by mass spectrometry.

However, molecular alterations occurring in small areas of the tissue or changes in the localization of the molecules within the tissue section are lost after homogenization, and smoothed away in retrieving the average molecular profile of the sample. Mass Spectrometry Imaging (MSI) aims at overcoming these limitations.

Mass Spectrometry Imaging entails the acquisition of one mass spectrum for each pixel of a digitalized tissue section [11]. In this way, local changes are fully preserved in place and not lost due to homogenization, and more representative molecular profiles can be extracted from regions of interest. After the MSI analysis, the spatial localization of the molecules of interest can be evaluated on-tissue by generating molecular images, which can be overlapped with histological cyto-morphological information obtained from stained tissue sections [11,12].

1.1.2 The MSI data

The data obtained after a MSI analysis is structured as a “data cube”, a tensor in which the two spatial dimensions (x and y axes) of the digitalized tissue section are combined with a third dimension consisting of the *mass-to-charge ratio* (*m/z*) of the molecules present within the tissue section [13].

For each *m/z* it is possible to display the spatial distribution / localization of the molecule(s) of interest on tissue by coloring the pixels according to the relative intensity of that *m/z* value in the corresponding spectra. A subset of the original data cube can be extracted, with spectra from specific Regions of Interest (ROIs).

1.1.3 Spectral preprocessing

Before submitting the MSI data to the statistical analysis, the spectral dataset must undergo a series of preprocessing steps, aimed at flattening the inter-sample and intra-sample pixel-by-pixel fluctuations in intensities due to sample preparation and mass spectrometric instrumentation [14]. In this way, the MS data is adequately prepared for statistical analysis, with a consequent enhancement of the biological information present within the data [15-18].

For instance, when on-tissue digestion is performed by spraying trypsin onto the sample, the enzyme activity varies throughout the tissue section, therefore releasing peptides in different amount and affecting the corresponding intensity in the spectrum [19]. In addition, matrix crystallization does not occur in the same manner at each pixel: this affects the analyte extraction process, which reflects onto the relative intensity of the corresponding *m/z* values in the mass spectrum. Finally, in the mass spectrometer, slight variations due to the electronic nature of the instrumentation can occur: ions can be generated in a different way depending on the crystallization of the matrix, the energy of the laser, the transmission efficiency of the analyzer, and the signal conversion by the digitizer [16,17].

Smoothing

Smoothing aims at discarding the fluctuations in the spectrum due mainly to the electronic nature of the mass spectrometer and impurities present within the sample [16]. The shape of the peaks results in being poorly defined and therefore the *peak picking* process can hardly discriminate the signal from the noise. The smoothing process enhances and eases the peak detection phase, since false positive peaks corresponding to electrical noise are discarded (Figures 2 and 3).

The smoothing can be performed by employing several algorithms:

- The *Savitzky-Golay filter* fits adjacent data points with a low-degree polynomial and takes the central point of the fitted polynomial curve as output; since it does not distort the essential features in the spectrum, this filter tends to fully preserve the intensity and the position of the peak.
- The *Moving Average* method defines a window which slides along the spectrum and replaces each data point with the average of the data points contained in the previous window.

- The *Gaussian* smoothing employs the Gaussian function to perform smoothing, the strength of which is determined by the σ parameter.

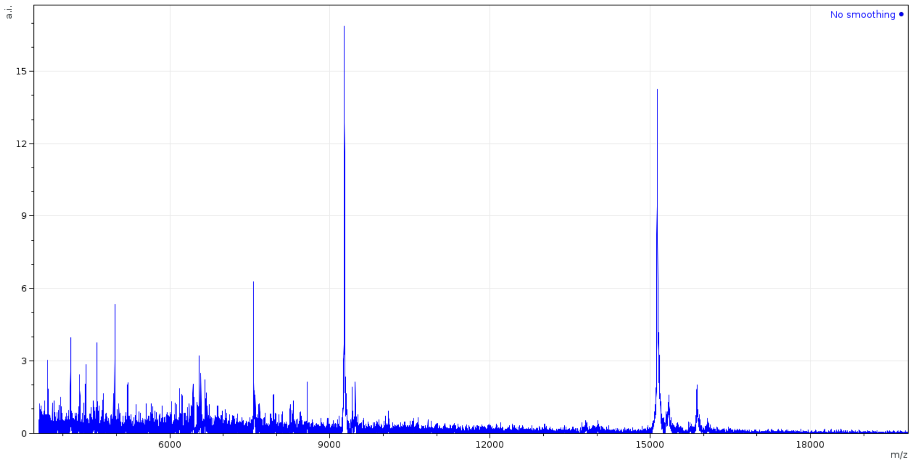


Figure 2: Example of a raw mass spectrum. The electrical noise present in the spectrum hinders the peak detection, by yielding false positive peaks coming from the noise.

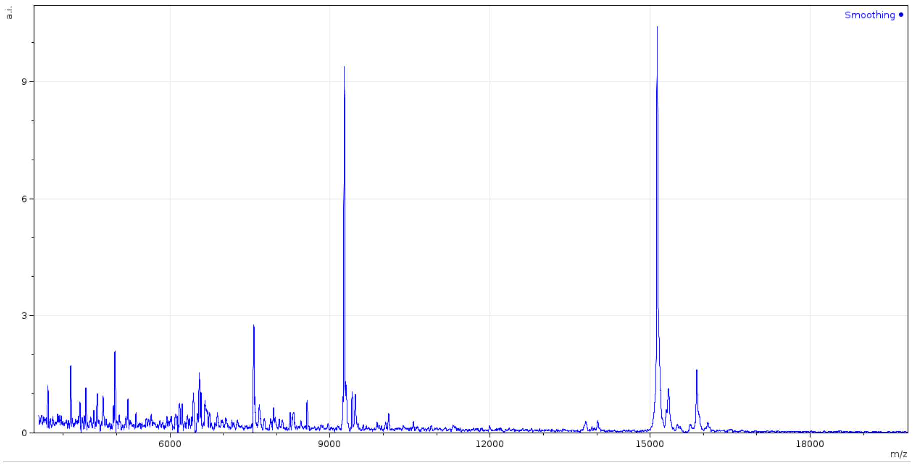


Figure 3: Example of a mass spectrum, after the application of a 21-point Savitzky-Golay smoothing filter. Peaks can now be more easily defined.

Baseline subtraction

The baseline of a spectrum is the line connecting the data points with lowest intensities, on which the entire spectrum lays [16]. A high baseline is essentially related to electrical noise and associated with chemical impurities in the sample. This hinders the estimation of the true intensity of a peak, compromising the fair comparison among spectra. The baseline subtraction process estimates and subtracts the baseline from the spectrum, by bringing the spectrum onto the x-axis, allowing for a more reliable estimation of the peak intensities and for more fair comparisons among peaks of different spectra (Figures 4 and 5).

Several algorithms can be employed for this task:

- The *TopHat* operator was designed for the extraction of small features from an image, based on the assumption that features of interest should stand out in a noisy environment. This algorithm computes *dilation* and *erosion* operations, in order to define the background of the spectrum, and finally it subtracts the result from the original spectrum.
- The *Convex Hull* method defines the baseline as a simple convex curve (a hull) connecting the two extremities of the spectrum. It is not a robust method, since it does not take into account local baseline variations, which can lead to the loss of informative portions of the mass spectrum.
- The *Median filter* estimates the points of the baseline curve by employing a moving median, in which a data point is the median of the previous window.
- The *Iterative Convolution* algorithm estimates the background of a spectrum by erasing the peaks from the spectrum by the use of a Gaussian filter, the amplitude of which can be defined as σ. The point-wise minimum of the background curve is taken as an estimation of the

baseline, which is then subtracted from the original spectrum.

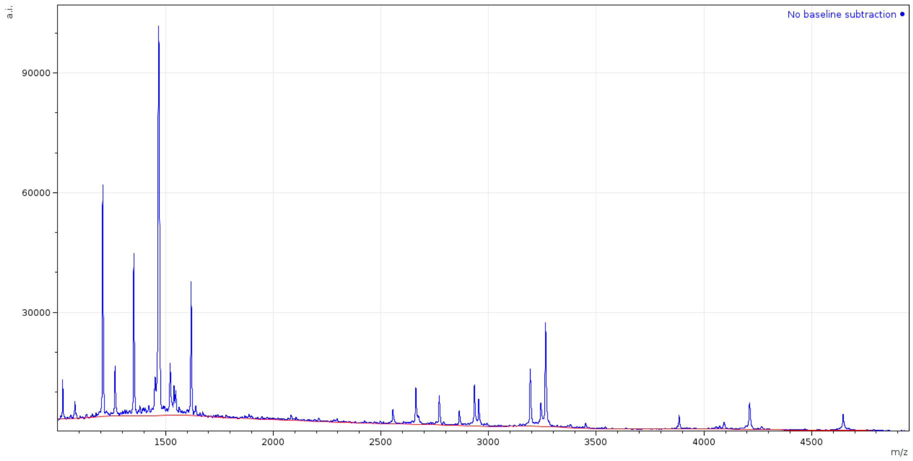


Figure 4: Example of baseline estimation on a raw mass spectrum.

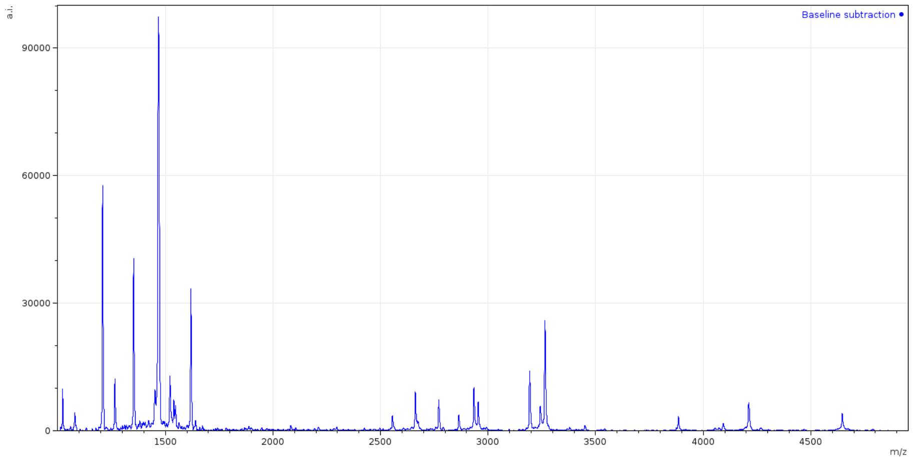


Figure 5: Example of a mass spectrum after undergoing baseline subtraction with the TopHat algorithm.

Normalization

The normalization of a mass spectrum consists in the multiplication of the absolute intensities in the mass spectrum by a scaling factor, which results in an intensity axis broadening or narrowing [16,18]. The intensity scale strongly depends upon the analog-digital conversion performed by the digitizer and the vendor of the instrument: for this reason, it is measured in “arbitrary units”. The aim of normalization, therefore, is to bring all the intensity values down to a common scale, in order to allow for more fair and vendor-free spectral comparisons.

- The *Total Ion Count* (TIC) method divides the spectrum intensities by the sum of all the intensity values for that spectrum (i.e. the total ion current), by generating spectra with the same area under the curve. The assumption behind the TIC normalization is that the spectra have a similar area and comparable number of peaks. Therefore, it is the most suitable normalization method for the majority of the MSI datasets.

  Artifacts can be generated when the assumptions are not satisfied, for example in the presence of a very intense ion in the spectrum (such as insulin for pancreas datasets): in this case, the TIC for that spectrum corresponds almost entirely to the intensity of that peak, resulting in a consequent suppression of all the other signals. In order to overcome this limitation, either a TIC with the exclusion of that peak or another algorithm (such as the *median*) can be used for normalization [18].

- The *Root Mean Square* (RMS) method divides the spectrum intensities by the square root of the sum of the intensity values for that spectrum squared. This method is mostly appropriate for use with datasets containing spectra that are expected to have small variations in the peak intensities. As for the TIC, this method can generate artifacts in the presence of prominent peaks [18].
- The *Median* method divides the spectrum intensities by the median intensity of that spectrum. This method has been found to be the most robust against the presence of very intense peaks within the mass spectrum, therefore more robust against peculiar conditions. However, the results obtained after median normalization depend on the type of noise in the spectra: if spectra do not contain a fully symmetrical

noise profile, this method will generate significant artifacts [18].

Peak picking

The peak picking extracts the information regarding the peaks present within the mass spectrum, along with the intensity of those peaks [16,20] (Figure 6). Different algorithms are employed for *peak picking* [21]: some of them estimate the spectrum noise and select only the peaks that extrude from the noise at a certain *signal-to-noise ratio* (*S/N*), other methods (such as the *Orthogonal Matching Pursuit*) estimate how much a peak resembles a Gaussian curve with a certain width [22,23].

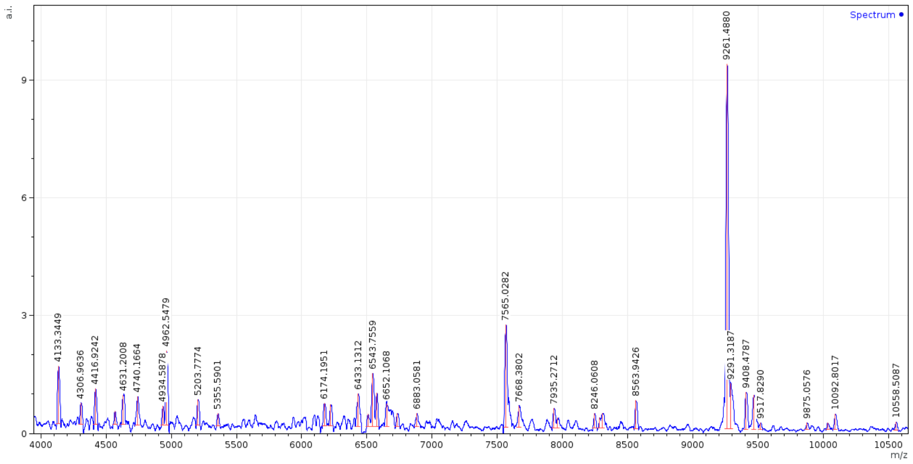


Figure 6: Example of a peak picking process performed on a mass spectrum. Only the information regarding the peak m/z and intensity is retained for the statistical analysis.

Finally, after peak maxima have been aligned to each other in order to account for fluctuations in the peak values among the spectra of the dataset related with the peak picking process, the data is ready to be submitted to the statistical analysis [22].

1.1.4 Machine learning and Mass Spectrometry Imaging

Machine learning is the branch of computer science aimed at the employment of algorithms that learn features from known data and make predictions onto new unknown data based upon its features [24-27]. Machine learning applications entail both unsupervised and supervised approaches, according to the data being unlabeled and labeled respectively, in order to discover and highlight hidden patterns within the data (through clustering) or to assign an unknown observation to a category (through classification and regression) [24-27].

Machine learning is applied in a great variety of fields nowadays and can be applied also to MSI data for clinical purposes [13].

Unsupervised learning

Unsupervised learning takes unlabeled data as input, i.e. data in which the outcome is not known. By the exploitation of the intrinsic information present within the data, clustering operations are performed in order to highlight hidden structures and/or patterns within the data, by estimating similarities among data observations, according to the pairwise distance among each other [28]. However, these approaches can be used also in a partially supervised manner, in such a way that the outcome of each observation is preserved during the unsupervised analysis but not taken into account by the algorithm, which performs its operations in blind, or by providing a defined number of clusters to obtain.

Hierarchical clustering analysis (HCA) is an agglomerative method that estimates the pairwise distance among data observations and generates a dendrogram (Figure 7), in which the observations are grouped together and placed under the same nodes (i.e. the joint point between clusters) according to the similarity among each other [29,30]. Being an agglomerative method, the algorithm starts by evaluating the individual data observations and grouping them in a bottom-up fashion, to generate clusters which, in turn, group with other clusters based upon their degree of similarity. Additionally, the dendrogram can be cut at a certain height, in order to generate a defined number of sub-clusters, which can resemble a particular situation under investigation.

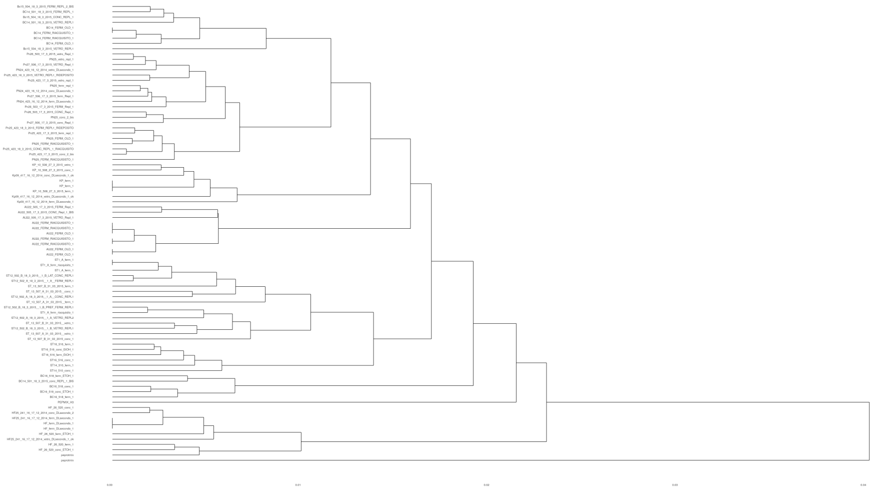


Figure 7: Example of a dendrogram.

In mass spectrometry imaging, data observations correspond to spectra, and spectra correspond to pixels: therefore, pixels corresponding to spectra under the same node can be colored in the same way, in such a way that a segmentation image is generated [31,32]. This unsupervised MS segmentation image, by resembling the dendrogram and allowing the visualization of spectral similarity on tissue, can highlight areas of interest on a molecular basis (Figure 8). Given that the analysis has been carried out in an unsupervised manner, the areas of interest highlighted by the HCA on-tissue visualization are depicted without the *a priori* knowledge regarding the presence of such areas in the tissue section. Therefore, this process allows the MS imaging approach to aid the diagnostic procedure by bringing areas of tissue to the attention of the pathologist and by highlighting the molecular changes even if not correlated with cyto-morphological features [31,32].

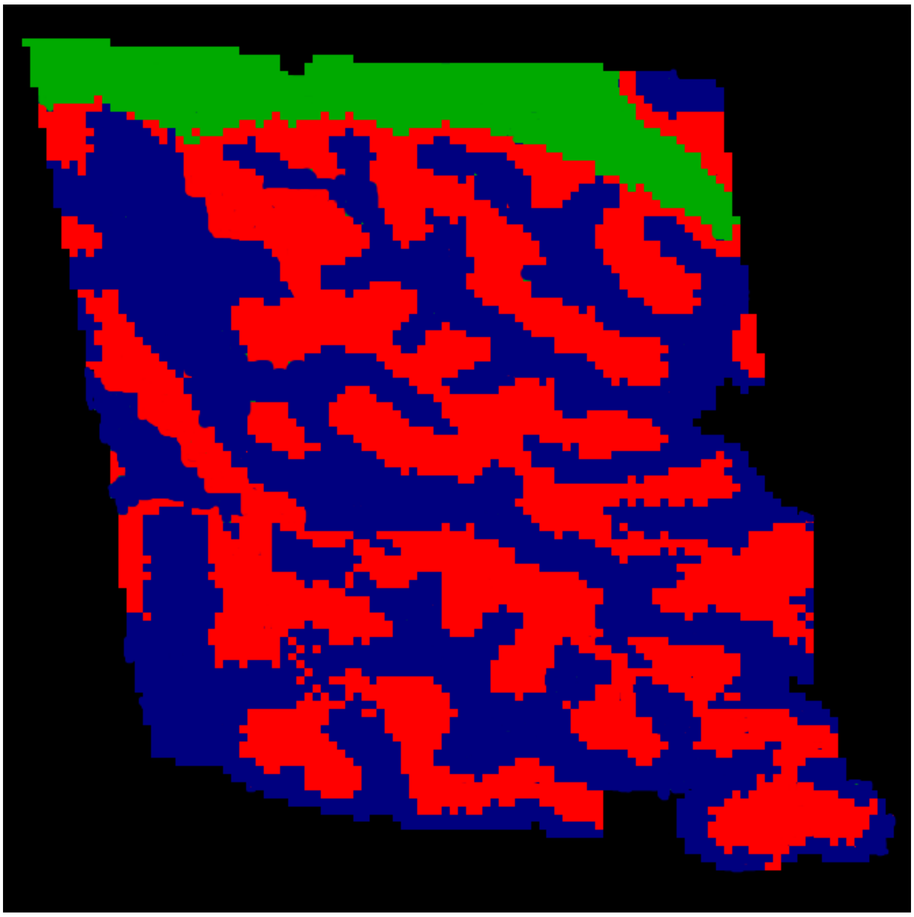


Figure 8: Example of an on-tissue hierarchical clustering [47]. Pixels corresponding to spectra under the same nodes are colored in the same way, generating a MS segmentation image which depicts tissue sub-areas according to the spectral similarity.

K-means clustering aims at partitioning the data observations into *k* clusters, in which each observation belongs to the cluster with the closest mean [33-35]: the algorithm finds the best centroids by iteratively assigning data points to clusters based upon the current centroids (points at the center of the cluster) and adjusting the new centroids (Figure 9).

The euclidean distance is used as metric and the variance is used as a measure of scatter: these two parameters may not be suitable for some type of data. The number of clusters *k* must be provided *a priori* for initializing the clustering operation: this implies that a number of expected clusters must be known and an inappropriately defined *k* may yield poor results. Since the clustering process relies on iterative optimization (heuristic algorithms), a convergence to a local minimum may produce non-definitive results, that may lead to misleading conclusions on the data; to overcome this limitation, however, it is possible to run it multiple times with different starting conditions.

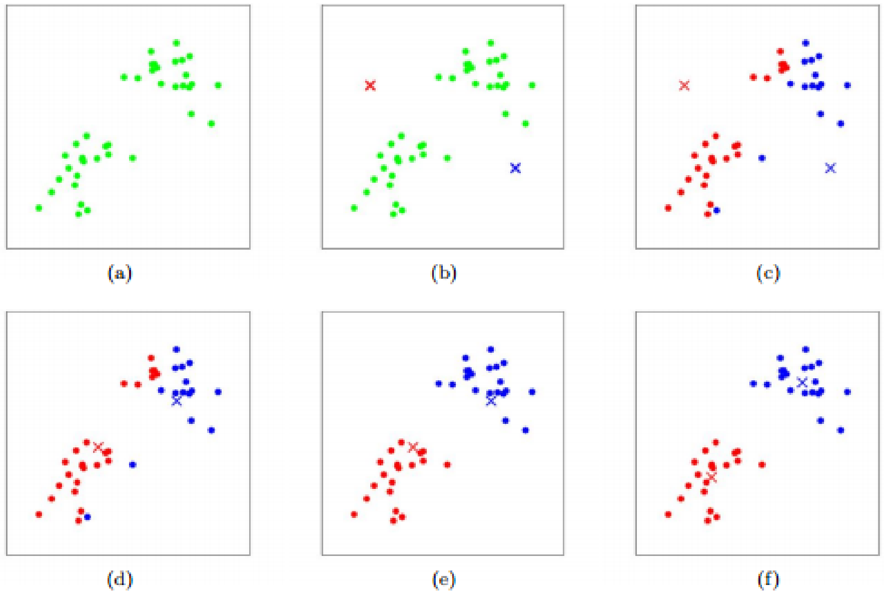


Figure 9: Graphical representation explaining the concept of the k-means clustering. The data observations of the initial dataset are placed in the feature space according to their value (a). Two symmetrical centroids are identified from the dataset (b), around which to group the closest observations (c). New more representative centroids are identified after the first clustering step (d) and observations are grouped again around them according to their distance to the centroids (e). New centroids are therefore identified, providing an adjustment to the previously identified centroids (f). The process iterates until it reaches convergence, i.e. the newly identified centroids are the same as the previous ones. [Picture taken from the Stanford University website]

As for the Hierarchical Clustering Analysis, in Mass Spectrometry Imaging it is possible to generate a segmentation image that resembles the clustering. Pixels corresponding to spectra within the same cluster are colored in the same way, identifying a defined (*k*) number of sub-regions on tissue, based upon spectral similarity (Figure 10).

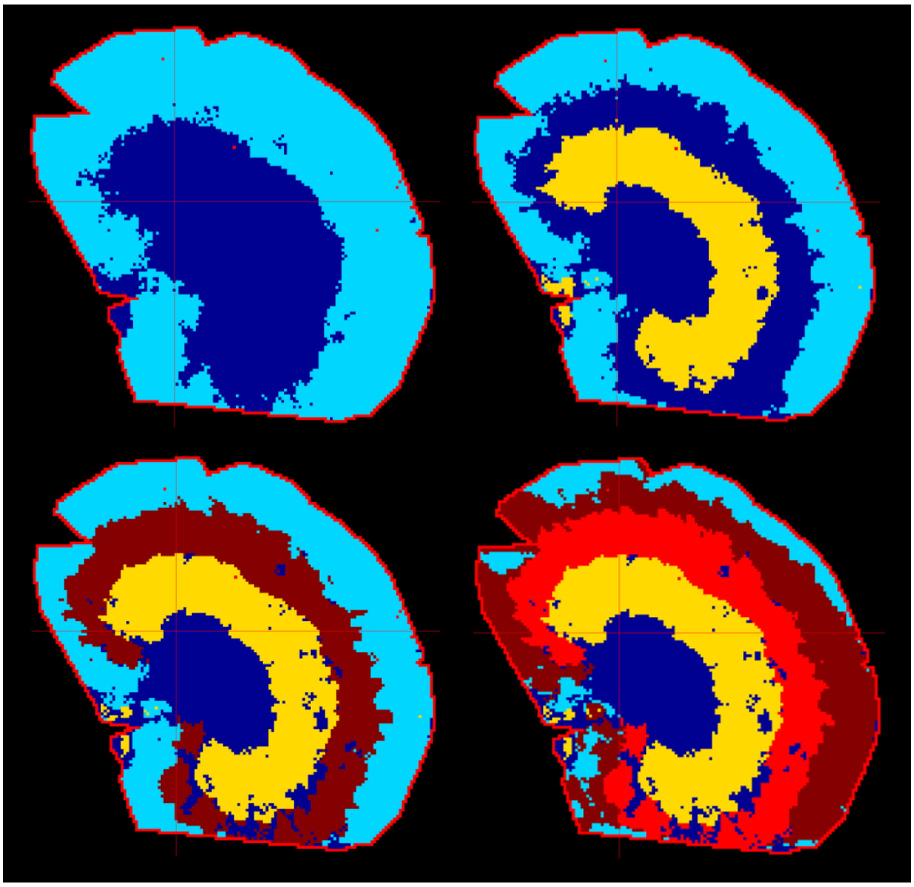


Figure 10: Example of an on-tissue k-means clustering, with a value of k set respectively to 2, 3, 4 and 5. Pixels corresponding to spectra grouped around the same centroid are colored in the same way, generating a MS segmentation image which depicts the expected (k) tissue sub-areas according to the spectral similarity.

Principal Component Analysis (PCA) aims at reducing the dimensionality of the data while preserving the information present within the data [36-38]. Given that MS datasets contain a number of features that ranges from hundreds to thousands of peaks, PCA provides an overview of the entire spectral dataset, by generating new variables (called *Principal Components*, PC) from the linear combination of the spectral features. The PCs are generated orthogonally to one another, in such a way that no redundancy among the new variables is present and from the first to the last PC a decreasing amount of variance is retained from the original dataset. In this way, by looking at the first Principal Components, an overview of almost the entire information present within the data can be obtained. The output of a PCA consists of a Score Chart (Figure 11) and a Loadings Plot (Figure 12). The former places data observations in a 2D or 3D graph according to the score of the PCs, allowing to evaluate the degree of similarity among the spectra by visualizing their distribution/clustering in the Principal Component space. The latter, by resembling the distribution of the former, allows to evaluate the most influencing feature which contribute more in driving the distribution/clustering of data observations in the Score Chart.

Despite being unsupervised, the PCA can be used also as a partially supervised method: in fact, it is possible to color the data points in the chart according to the outcome, in order to evaluate if the data has intrinsic properties to discriminate between different conditions/classes.

In this way, by the combination of the two plots, not only it is possible to determine if the data is capable of discriminating among different classes, but also putative signals of interest can be highlighted for further investigation [13].

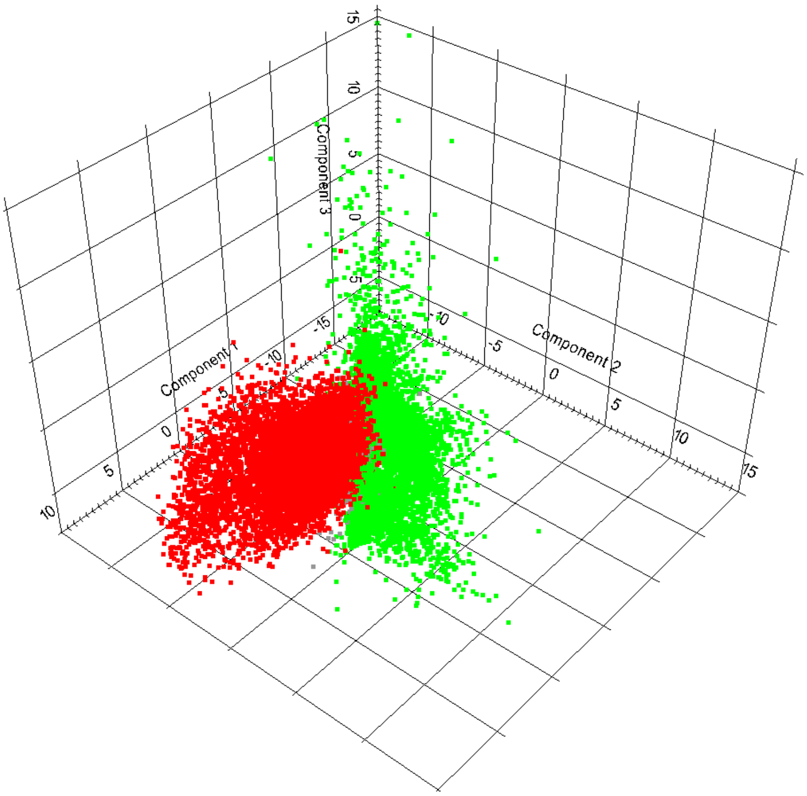


Figure 11: Example of a PCA score chart, in which each point corresponds to a spectrum in the MSI dataset. In this example, PCA is used as a partially supervised data mining approach: two classes are loaded and displayed in two different colors. The capability of the data to discriminate between the two classes can be assessed.

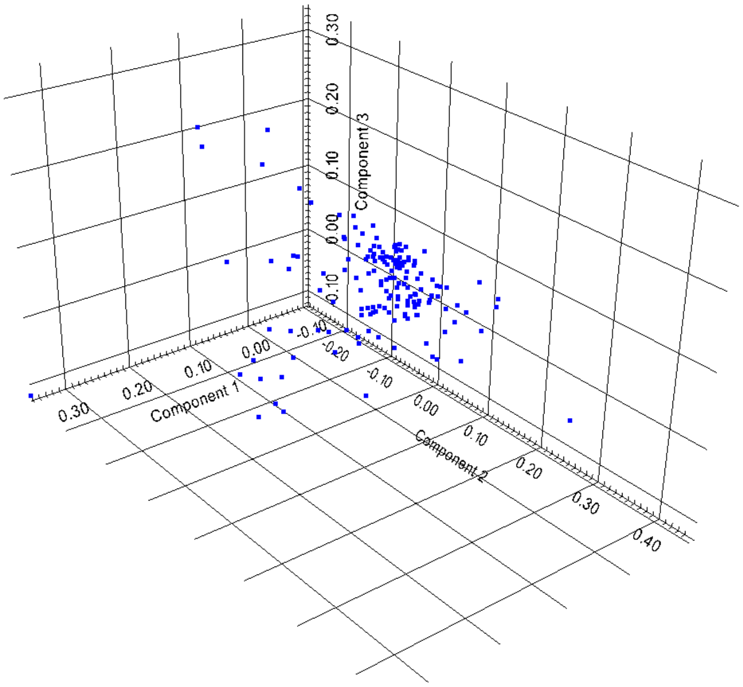


Figure 12: Example of a PCA loadings plot, in which each point corresponds to a spectral feature (m/z). The most influencing peaks can be addressed as the most isolated points in the plot.

In Mass Spectrometry Imaging, the results of a PCA can be translated into a MS segmentation image, which resembles the distribution of the data points according to the Principal Components. The pixels, in fact, are colored according to the score of the Principal Component computed onto the corresponding spectrum (Figure 13). Therefore, sub-areas of interest can be visualized on-tissue based upon spectral similarity only, since Principal Components with similar score denote similar spectra.

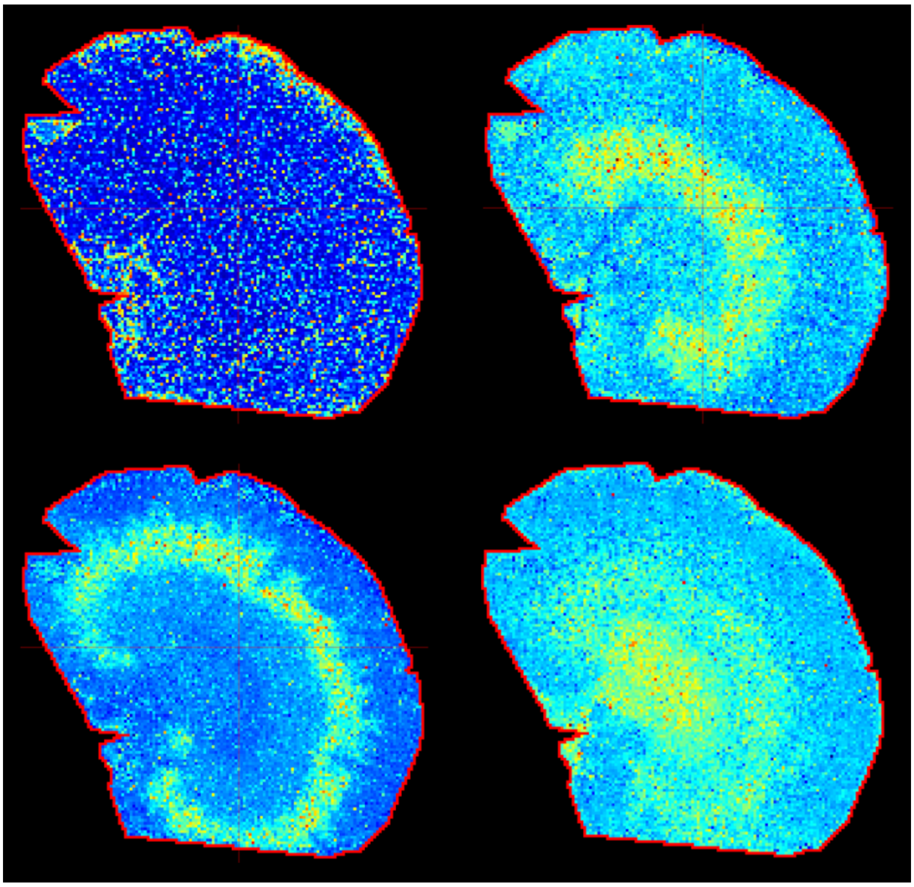


Figure 13: Example of an on-tissue PCA, in which each pixel is colored according to the score of the Principal Component associated with the corresponding spectrum. In this figure, the results of the PCA analysis are displayed on-tissue: the scores of the PC 1, 2, 3 and 4 are plotted.

Supervised learning

Supervised methods aims at learning from labeled data, i.e. data in which the outcome is known, in order to exploit known features to make predictions on new unknown data [39,40] (Figure 14). This resembles, mostly, the classification problem, in which algorithms, referred to as classifiers, learn from features provided by labeled data in order to predict the outcome of unknown observations based upon their features (which correspond to *m/z* peaks in the mass spectrometry imaging dataset) [39,40]. When a large number of features is present, a feature selection step may be required in order to make classifiers faster and more reliable [41-43].

Regression is another widely employed supervised statistical approach, in which the response variable is not discrete, but continuous [39, 40]. In a similar way of what happens for classification, in regression, models predict the value of an outcome variable based upon features that have been learned in the training phase.

Feature selection Feature selection (also referred to as variable selection or attribute selection) is the process of retaining only the most informative features, by discarding the least relevant ones, which are either invariant or provide redundant information within the dataset [41-43]. When classifier algorithms are trained with the employment of all the features of the training dataset, the risk that arises is the non-translatability of the trained classification system on other datasets: in fact, the algorithm, not only results in being more complex in its construction, but, by relying upon all the features of the training set for computing classification, becomes highly specific for the training data and results incapable of performing classification on other unknown datasets. In this case, the classifier suffers from overfitting, by excessively fitting the training data. Feature selection, therefore, aims at reducing the complexity of trained classifiers and their tendency of suffering from overfitting, and at speeding up the computations by alleviating the overhead provided by possible confounding factors within the features. Therefore, the issues related with the “curse of dimensionality” (related to the presence of a high amount of features, i.e. dimensions, which leads to the requirement of many observations in the training dataset to account for the possible presence of the highest amount of combinations of feature values possible) are overcome and the results provided by the selected subset of informative features are subjected to a translatable interpretation by researchers. There are three types of feature selection methods [42]: *filter methods* are fast and employ the use of a proxy measure to select the most influencing features, however by not taking into account the possible relationship between features; *wrapper methods* employ predictive models to evaluate the impact of the features onto the classification which is directly correlated with the variable importance and takes into account possible relationships between features; *embedded methods* perform feature selection while constructing the predictive model.

Classification problem In the classification problem paradigm, the first phase, the training phase, allows classifiers to build the mathematical formula by taking labeled data as input, to discriminate with different techniques among the provided categories. For example, *Support Vector Machines* fit a hyperplane, with the aid also of kernel functions, aiming at maximizing the sum of the absolute distances between the separating hyperplane itself and the closest points belonging to each class (i.e. the so-called margin) [44]. *K-Nearest Neighbors* establish the category an observation belongs to by calculating the distance between the observations and evaluating the most frequent category among the *k* closest observations [45]. *Random Forests* build a multitude of decision trees, in each of which thresholds of feature values determine whether the observation belongs to a class or to another: each tree predicts the outcome of an observation and the result is the mode class of all predictions [46].

In the validation phase, the hyper-parameters of the model (such as the trade-off parameter C in a SVM) undergo tuning, in order to estimate the best combination for maximizing the classification capability of the model itself.

Finally, in the test phase, the classifier performances are evaluated by the predictions made onto partitions of the same training set (cross-validation) or onto an external labeled dataset (external validation). In (*k*-fold) cross-validation, the training set is subdivided into *k* subsets, of approximately the same size: at each iteration, the learning process is repeated by employing one subset for testing and the other subsets for training; after k iterations, the results are average to get a single estimation of the behavior of the classifier. In the external validation, an independent validation set is given to the classifier, to assess its capability of correctly classifying unknown data. The discrepancy between the predicted class and the actual class gives the performance parameters of the model, such as sensitivity (True Positive Rate, TPR, i.e. the proportion of positive subjects that are really positive), specificity (True Negative Rate, TNR, i.e. the proportion of negative subjects that are really negative), Positive Predictive Value (PPV, i.e. the proportion of positive test results that are associated with really positive subjects) and Negative Predictive Value (NPV, i.e. the proportion of negative test results that are associated with really negative subjects). This phase denotes the generalization capability of the inferred model, i.e. the capability of the model in being applied to a new dataset.

At the end of the process, the classifier can be employed for making predictions on new data, which can be also weighed according to the performance parameters evaluated in the previous phases.

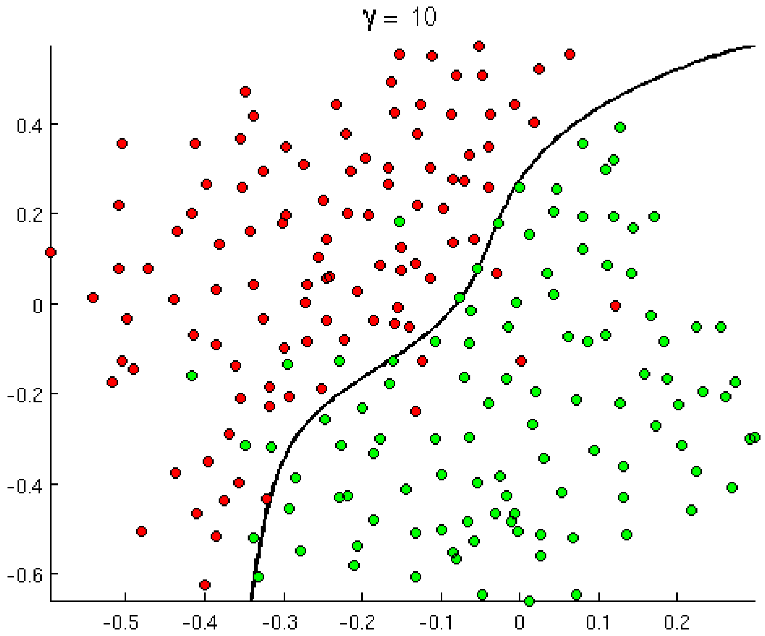


Figure 14: Graphical representation of the classification problem concept. A classifier algorithm defines a mathematical formula that links the values of the features of the data points to their outcome, in order to exploit such features for future predictions on unlabeled data. The figure represents an example Support Vector Machine classifier: features are centered and scaled (subtracted by the mean and divided by the standard deviation) and data observations are placed in the feature space according to the values of the features (represented onto the x and y axes). The color represents the belonging class of the data observations (green for benign and red for malignant patients), while the black curve the separating hyperplane of the Support Vector Machine classifier. Gamma represents a hyper-parameter of the SVM, i.e. the influence of a single training example [Picture taken from the Stanford University website]

In Mass Spectrometry Imaging, an on-tissue classification can be obtained by generating a MS segmentation image resembling the classification by coloring pixels according to the predicted class (Figure 15) (Chapters 3 and 4).

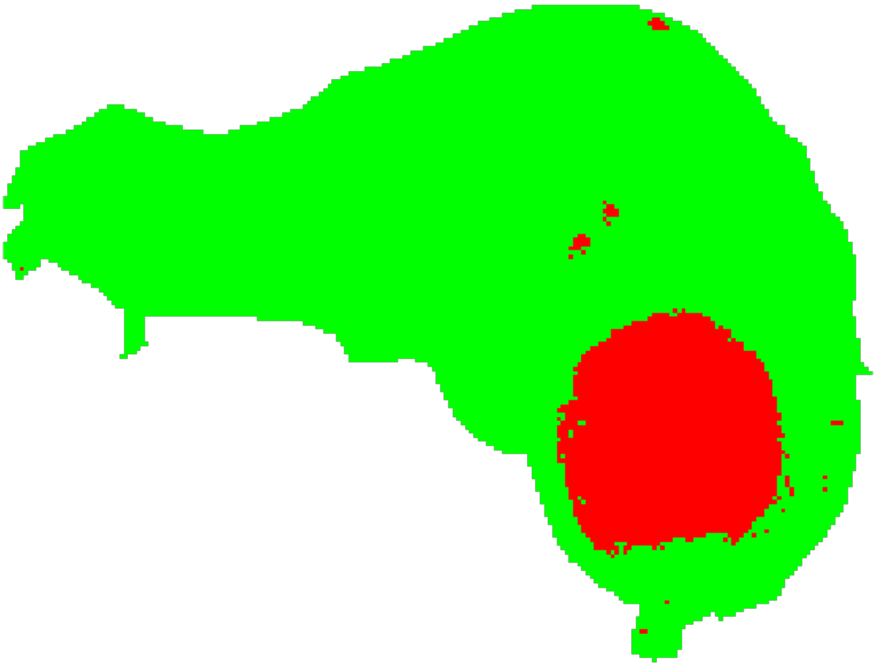


Figure 15: Example of an on-tissue classification, in which each pixel is colored according to the predicted class of the corresponding spectrum. In this example, red pixels correspond to malignant spectra, while green pixels to benign spectra (Chapter 4).

1.2 Scope of the thesis

The work performed during this three-year PhD project has been focused on the development of an easy-to-use software tool for the statistical analysis of clinical Mass Spectrometry Imaging (MSI) data, in particular for patient classification, aimed at aiding the diagnostic procedure in the daily clinical practice.

Chapter 2: evaluation of the state-of-the art application of MALDI-MSI in the clinical environment, with particular attention towards spectral preprocessing, machine learning algorithms (both unsupervised and supervised learning) and software implementation.

Chapter 3: a first application of state-of-the-art machine learning classification algorithm (SVM), coupled with a wrapper feature selection method (RFE), for the classification of thyroid cytological smears via MALDI-MSI; first proposal of the pixel-by-pixel classification concept.

Chapter 4: development of a full software program (wrapped under a simple and intuitive graphical user interface) for the generation and application of an ensemble classification system, which combines different classifiers (with different weights) for the pixel-by-pixel classification of biopsies via MALDI-MSI.

Chapter 5: development of an application, with a simple and intuitive graphical user interface, for the generation of G-code method files for the iMatrixSpray device, for the preparation of samples for the MALDI-MSI analysis.

References

[1] E. De Hoffmann and V. Stroobant, *Mass Spectrometry - Principles and Applications.*, vol. 29, no. 6. 2007.

[2] J. B. Fenn, “Electrospray wings for molecular elephants (Nobel lecture),” in *Angewandte Chemie - International Edition*, 2003, vol. 42, no. 33, pp. 3871–3894.

[3] J. Peter-Katalinić and F. Hillenkamp, *MALDI MS: A Practical Guide to Instrumentation, Methods and Applications*. 2007.

[4] D. S. McPhail, “Applications of Secondary Ion Mass Spectrometry (SIMS) in materials science,” *J. Mater. Sci.*, vol. 41, no. 3, pp. 873–903, 2006.

[5] J. M. Wiseman and B. C. Laughlin, “Desorption Electrospray Ionization (DESI) Mass Spectrometry: A brief introduction and overview,” *Bioanal. Syst.*, vol. 22, no. 1, pp. 11–14, 2007.

[6] R. E. March, “Quadrupole ion traps,” *Mass Spectrom. Rev.*, vol. 28, no. 6, pp. 961–989, 2009.

[7] P. S. H. Wong, “Ion Trap Mass Spectrometry,” *Rapid Commun. Mass Spectrom.*, vol. 54, no. 1, pp. 237–76, 1997.

[8] R. A. Zubarev and A. Makarov, “Orbitrap mass spectrometry,” *Anal. Chem.*, vol. 85, no. 11, pp. 5288–5296, 2013.

[9] E. N. Nikolaev, “Some notes about FT ICR mass spectrometry,” *Int. J. Mass Spectrom.*, vol. 377, no. 1, pp. 421–431, 2015.

[10] H. Wollnik, “Time-of-flight mass analyzers,” *Mass Spectrom. Rev.*, vol. 12, no. 2, pp. 89–114, 1993.

[11] R. M. Caprioli, T. B. Farmer, and J. Gile, “Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS.,” *Anal. Chem.*, vol. 69, no. 23, pp. 4751–60, 1997.

[12] R. Casadonte *et al.*, “Imaging mass spectrometry to discriminate breast from pancreatic cancer metastasis in formalin-fixed paraffin-embedded tissues,” *Proteomics*, vol. 14, no. 7–8, pp. 956–964, 2014.

[13] M. Galli, I. Zoppis, A. Smith, F. Magni, and G. Mauri, “Machine learning approaches in MALDI-MSI: clinical applications,” *Expert Rev. Proteomics*, vol. 13, no. 7, pp. 685–696, 2016.

[14] P. Ràfols *et al.*, “Signal preprocessing, multivariate analysis and software tools for MA(LDI)-TOF mass spectrometry imaging for biological applications,” *Mass Spectrometry Reviews*, 2016.

[15] E. a. Jones, S.-O. Deininger, P. C. W. Hogendoorn, A. M. Deelder, and L. a. McDonnell, “Imaging mass spectrometry statistical analysis,” *J. Proteomics*, vol. 75, no. 16, pp. 4962–4989, 2012.

[16] E. H. Jeremy L. Norris, Dale S. Cornett, James A. Mobley, Malin Andersson and and R. M. C. Seeley, Pierre Chaurand, “Processing MALDI Mass Spectra to Improve Mass Spectral Direct Tissue Analysis,” *Int J Mass Spectrom.*, vol. 260, no. 2–3, pp. 212–221, 2007.

[17] R. J. A. Goodwin, “Sample preparation for mass spectrometry imaging: Small mistakes can lead to big consequences,” *J. Proteomics*, vol. 75, no. 16, pp. 4893–4911, 2012.

[18] S. O. Deininger *et al.*, “Normalization in MALDI-TOF imaging datasets of proteins: Practical considerations,” *Anal. Bioanal. Chem.*, vol. 401, no. 1, pp. 167–181, 2011.

[19] B. Cillero-Pastor and R. M. a Heeren, “Matrix-Assisted Laser Desorption Ionization Mass Spectrometry Imaging for Peptide and Protein Analyses: A Critical Review of On-Tissue Digestion.,” *J. Proteome Res.*, 2013.

[20] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature Extraction: Foundations and Applications*, vol. 207. 2006.

[21] C. Yang, Z. He, and W. Yu, “Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis.,” *BMC Bioinformatics*, vol. 10, p. 4, 2009.

[22] D. Trede, J. H. Kobarg, J. Oetjen, H. Thiele, P. Maass, and T. Alexandrov, “On the importance of mathematical methods for analysis of MALDI-imaging mass spectrometry data.,” *J. Integr. Bioinform.*, vol. 9, no. 1, p. 189, 2012.

[23] T. Alexandrov, “MALDI imaging mass spectrometry: statistical data analysis and current computational challenges.,” *BMC Bioinformatics*, vol. 13 Suppl 1, no. Suppl 16, p. S11, 2012.

[24] T. M. Mitchell, *Machine Learning*, Internatio., no. 1. McGraw-Hill Education, 1997.

[25] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. 2000.

[26] B. C.M., “Pattern Recognition and Machine Learning,” *Springer*, 2006.

[27] D. De Ridder, J. De Ridder, and M. J. T. Reinders, “Pattern recognition in bioinformatics,” *Brief. Bioinform.*, vol. 14, no. 5, pp. 633–647, 2013.

[28] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data.pdf*, vol. 355. 1988.

[29] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Hierarchical Clustering*. 2011.

[30] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering: An overview,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 1, pp. 86–97, 2012.

[31] T. Alexandrov *et al.*, “Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering,” *J. Proteome Res.*, vol. 9, no. 12, pp. 6535–6546, 2010.

[32] T. Alexandrov and J. H. Kobarg, “Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering,” *Bioinformatics*, vol. 27, no. 13, pp. 230–238, 2011.

[33] J. A. Hartigan and M. A. Wong, “A K-Means Clustering Algorithm,” *Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979.

[34] A. W. Moore, “K-means and Hierarchical Clustering,” *Stat. Data Min. Tutorials*, pp. 1–24, 2001.

[35] D. T. Pham, S. S. Dimov, and C. D. Nguyen, “Selection of K in K-means clustering,” *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.*, vol. 219, no. 1, pp. 103–119, 2005.

[36] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4. pp. 433–459, 2010.

[37] I. T. Jolliffe, *Principal Component Analysis, Second Edition*, vol. 30, no. 3. 2002.

[38] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, 2016.

[39] S. B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques,” *Informatica*, vol. 31, pp. 249–268, 2007.

[40] MathWorks, S. M. Chelly, C. Denis, and MathWorks, “Applying Supervised Learning,” *What is Mach. Learn.*, vol. 33, no. 2, pp. 326–333, 2016.

[41] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 1157–1182, 2003.

[42] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19. pp. 2507–2517, 2007.

[43] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[44] N. Cristianini and J. Shawe-Taylor, *An introduction to Support Vector Machines*, vol. 47, no. 2. Cambridge, UK: Cambridge University Press, 2000.

[45] L. E. Peterson, “K-nearest neighbor,” *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[46] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[47] G. De Sio *et al.*, “A MALDI-Mass Spectrometry Imaging method applicable to different formalin-fixed paraffin-embedded human tissues,” *Mol. BioSyst.*, vol. 11, no. 6, pp. 1507–1514, 2015.

Chapter 2

Machine learning approaches in MALDI-MSI: clinical applications

Manuel Galli[1,*], Italo Zoppis[2,*], Andrew Smith[1], Fulvio Magni[1], and Giancarlo Mauri[2]

[1]Department of Medicine and Surgery, University of Milano-Bicocca, Monza Brianza, Italy ;

[2]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano, Italy

**These authors contributed equally to this work*

Expert Review of Proteomics

June 23 2016

doi: 10.1080/14789450.2016.1200470

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.

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 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 statistical 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 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 (i.e. with equal variances among different classes/groups) 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) (the p-norm of a vector is defined as the 1/p root of the sum of all the elements of the vector at the power of p): 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 pre-processing 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.

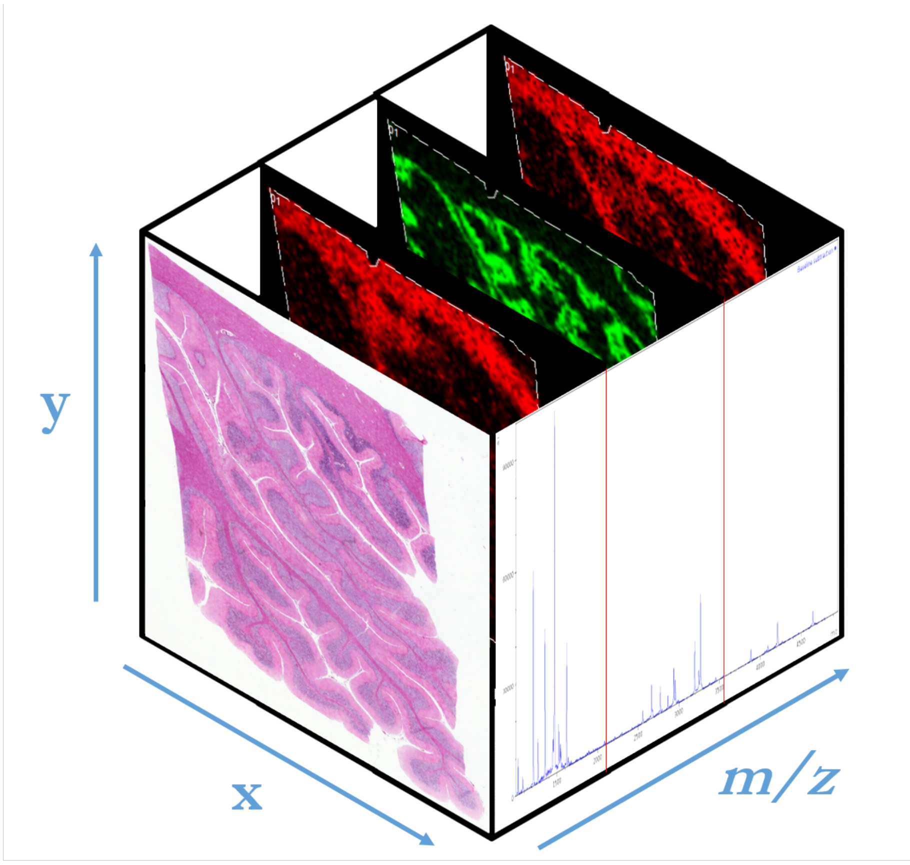


Figure 1: MALDI-MSI data cube. The x and y axes represent the spatial coordinates of the 2D digitalized tissue image (a human cerebellum tissue section is shown as an example); the z-axis represents the mass-to-charge (m/z) values in the acquired spectra. For each m/z value in the spectrum, a 2D molecular 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.

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

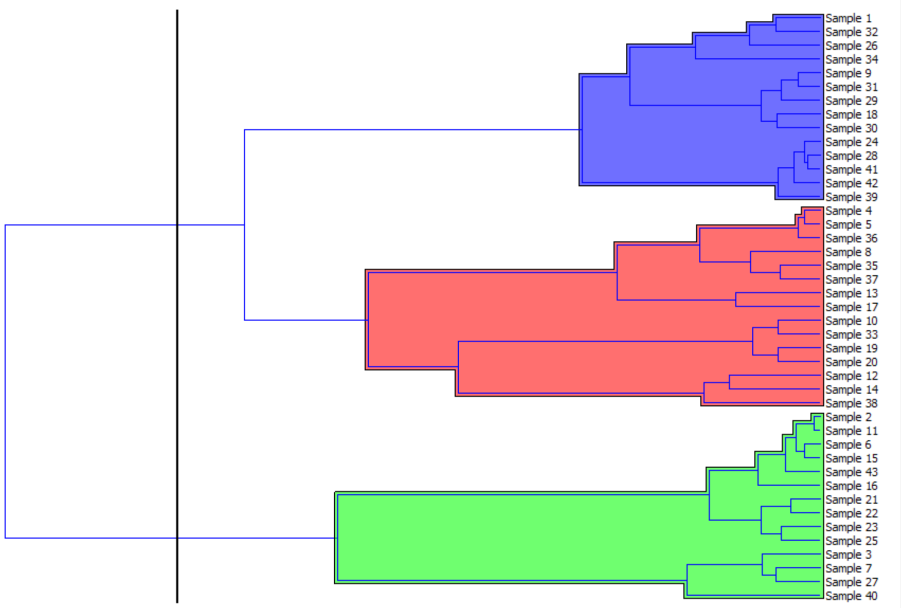


Figure 2: The figure displays an example dendrogram, in which average profiles of samples are clustered together based upon similarity only. The software (Orange Canvas in this instance) calculates the distance between samples and places similar profiles under the same node. A defined number of clusters 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 an 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.

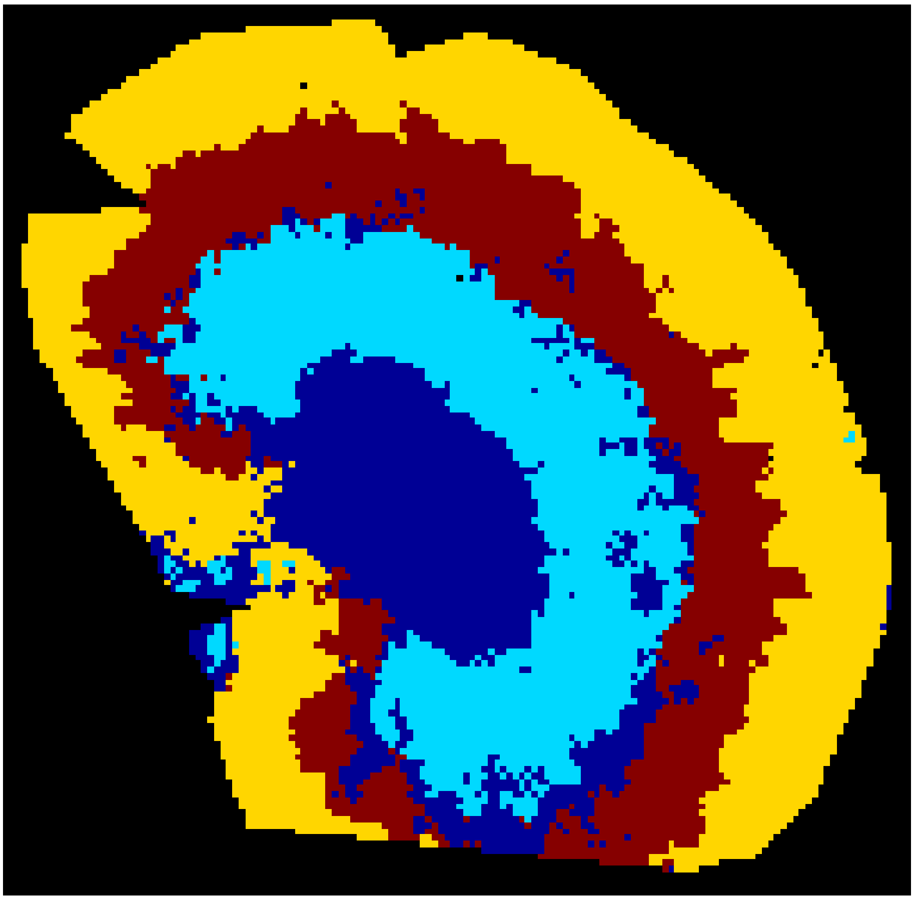


Figure 3: The figure displays an example segmentation map, obtained with SCiLS Lab 2014, onto a section of rat kidney, by coloring pixels corresponding to spectra under the same nodes with the same color. This approach has been able to correctly identify tissue sub-areas, perfectly overlapping with histo-morphological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

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 sub-areas of gastric cancer and human cerebellum. MALDI-MSI can highlight tissue sub-areas 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 sub-structures (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 re-evaluation 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 sub-populations. This has a deep impact onto the clinical outcome of a treatment, since a drug can be effective on some sub-populations 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 sub-populations 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 sub-areas within the same tumor region, that are histologically homogeneous [52]. Therefore, intratumor heterogeneity can be investigated, to the point where molecular signatures for each sub-population 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 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 sub-populations [56]. After the identification of myxofibrosarcoma tumor sub-populations, 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 non-informative 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 non-parametric 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 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 semistructured 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 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 sub-areas 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.

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.

7. Trim PJ, Snel MF. Small molecule MALDI MS imaging: current technologies and future challenges. Methods. 2016. pii:S1046-2023(16)30011-1.

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. 1st 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 intratumour 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.

Supplementary material

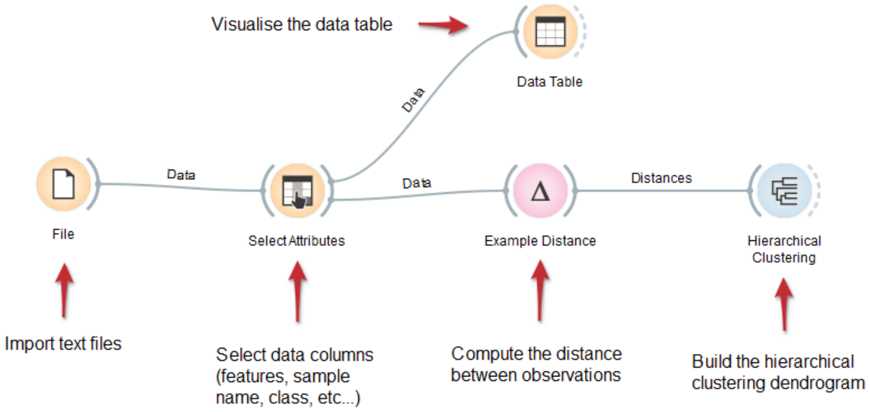
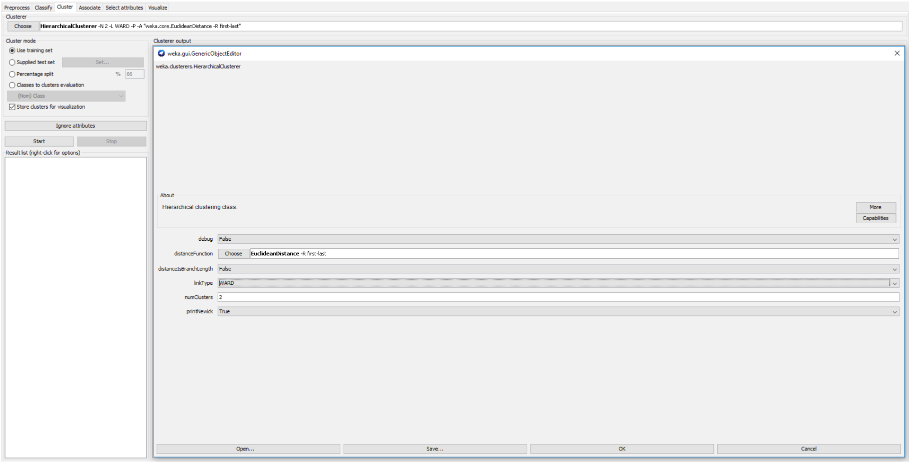

Visualise the data table
Data Table
Data
File
Data
Select Attributes
Data
Example Distance
Distances
Hierarchical Clustering
Import text files
Select data columns (features, sample name, class, etc...)
Compute the distance between observations
Build the hierarchical clustering dendrogram

Figure S1: The figure displays a typical workflow for cluster analysis in Orange Canvas. After the data matrix has been imported, the software allows to select the data columns to be considered as attributes, class and meta-attributes, in order to include or exclude data columns in the analysis. The distance between data rows (observations) is computed in the "Example Distance" widget, before running the hierarchical clustering process.



Clusterer: HierarchicalClusterer -N 2 -L WARD -P -A "weka.core.EuclideanDistance -R first-last"

Cluster mode
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) Class
☑ Store clusters for visualization

Result list (right-click for options)

Clusterer output

weka.gui.GenericObjectEditor

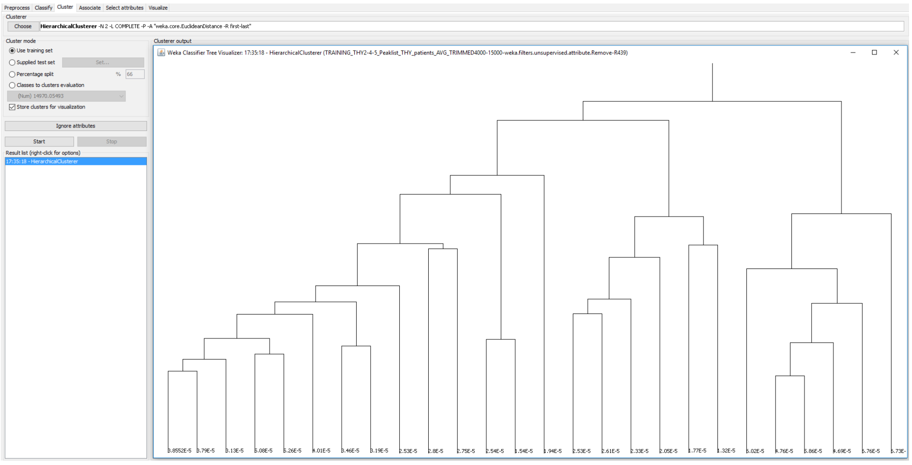
weka.clusterers.HierarchicalClusterer

About

Hierarchical clustering class.

debug: False
distanceFunction: EuclideanDistance -R first-last
distanceIsBranchLength: False
linkType: WARD
numClusters: 2
printNewick: True

Figure S2: The figure displays a typical workflow for clustering in Weka. The clustering algorithm can be chosen in the "Clusterer" entry, while the parameters can be tuned by clicking onto the entry itself (the figure shows the window with the algorithm parameters).



Clusterer: HierarchicalClusterer -N 2 -L COMPLETE -P -A "weka.core.EuclideanDistance -R first-last"

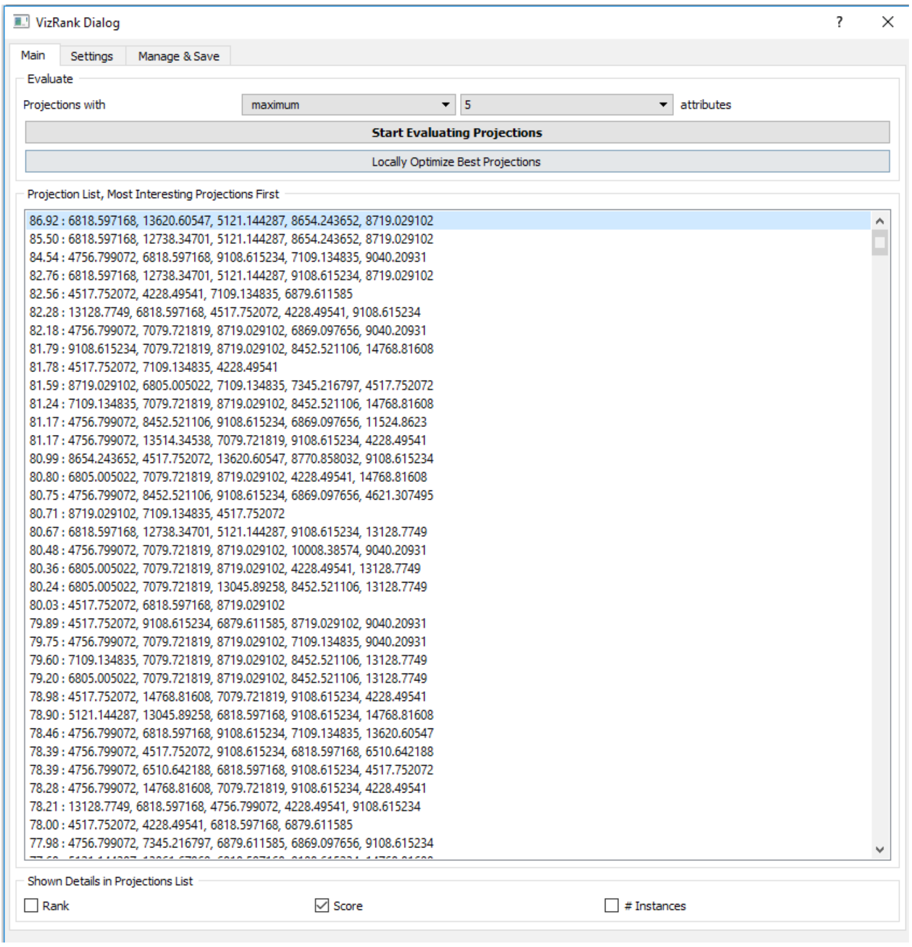
Cluster mode
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) 14970.05493
☑ Store clusters for visualization

Result list (right-click for options)
17:35:18 - HierarchicalClusterer

Clusterer output

Weka Classifier Tree Visualizer: 17:35:18 - HierarchicalClusterer (TRAINING_THY2-4-5_Peaklist_THY_patients_AVG_TRIMMED4000-15000-weka.filters.unsupervised.attribute.Remove-R439)

Figure S3: The figure displays a dendrogram returned by Weka after running the hierarchical clustering analysis.


VizRank Dialog
?
Main
Settings
Manage & Save
Evaluate
Start Evaluating Projections
Locally Optimize Best Projections

Projections with maximum 5 attributes

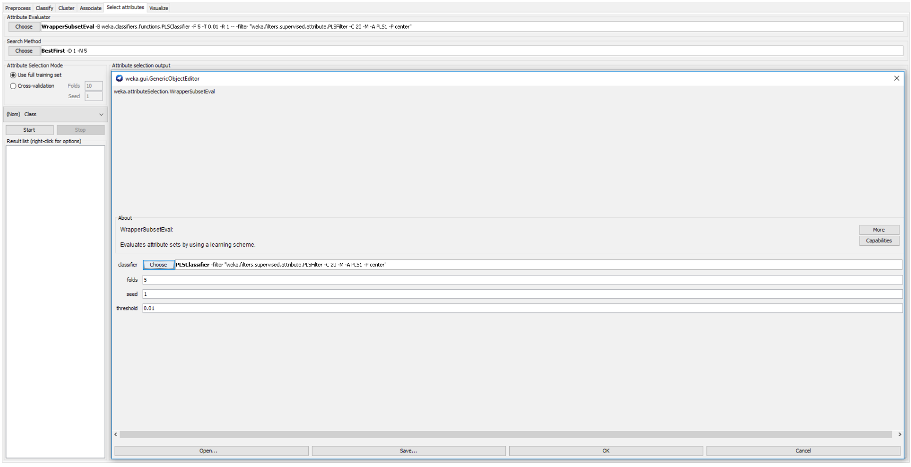
Projection List, Most Interesting Projections First

```
86.92 : 6818.597168, 13620.60547, 5121.144287, 8654.243652, 8719.029102
85.50 : 6818.597168, 12738.34701, 5121.144287, 8654.243652, 8719.029102
84.54 : 4756.799072, 6818.597168, 9108.615234, 7109.134835, 9040.20931
82.76 : 6818.597168, 12738.34701, 5121.144287, 9108.615234, 8719.029102
82.56 : 4517.752072, 4228.49541, 7109.134835, 6879.611585
82.28 : 13128.7749, 6818.597168, 4517.752072, 4228.49541, 9108.615234
82.18 : 4756.799072, 7079.721819, 8719.029102, 6869.097656, 9040.20931
81.79 : 9108.615234, 7079.721819, 8719.029102, 8452.521106, 14768.81608
81.78 : 4517.752072, 7109.134835, 4228.49541
81.59 : 8719.029102, 6805.005022, 7109.134835, 7345.216797, 4517.752072
81.24 : 7109.134835, 7079.721819, 8719.029102, 8452.521106, 14768.81608
81.17 : 4756.799072, 8452.521106, 9108.615234, 6869.097656, 11524.8623
81.17 : 4756.799072, 13514.34538, 7079.721819, 9108.615234, 4228.49541
80.99 : 8654.243652, 4517.752072, 13620.60547, 8770.858032, 9108.615234
80.80 : 6805.005022, 7079.721819, 8719.029102, 4228.49541, 14768.81608
80.75 : 4756.799072, 8452.521106, 9108.615234, 6869.097656, 4621.307495
80.71 : 8719.029102, 7109.134835, 4517.752072
80.67 : 6818.597168, 12738.34701, 5121.144287, 9108.615234, 13128.7749
80.48 : 4756.799072, 7079.721819, 8719.029102, 10008.38574, 9040.20931
80.36 : 6805.005022, 7079.721819, 8719.029102, 4228.49541, 13128.7749
80.24 : 6805.005022, 7079.721819, 13045.89258, 8452.521106, 13128.7749
80.03 : 4517.752072, 6818.597168, 8719.029102
79.89 : 4517.752072, 9108.615234, 6879.611585, 8719.029102, 9040.20931
79.75 : 4756.799072, 7079.721819, 8719.029102, 7109.134835, 9040.20931
79.60 : 7109.134835, 7079.721819, 8719.029102, 8452.521106, 13128.7749
79.20 : 6805.005022, 7079.721819, 8719.029102, 8452.521106, 13128.7749
78.98 : 4517.752072, 14768.81608, 7079.721819, 9108.615234, 4228.49541
78.90 : 5121.144287, 13045.89258, 6818.597168, 9108.615234, 14768.81608
78.46 : 4756.799072, 6818.597168, 9108.615234, 7109.134835, 13620.60547
78.39 : 4756.799072, 4517.752072, 9108.615234, 6818.597168, 6510.642188
78.39 : 4756.799072, 6510.642188, 6818.597168, 9108.615234, 4517.752072
78.28 : 4756.799072, 14768.81608, 7079.721819, 9108.615234, 4228.49541
78.21 : 13128.7749, 6818.597168, 4756.799072, 4228.49541, 9108.615234
78.00 : 4517.752072, 4228.49541, 6818.597168, 6879.611585
77.98 : 4756.799072, 7345.216797, 6879.611585, 6869.097656, 9108.615234
```

Shown Details in Projections List

□ Rank ☑ Score □ # Instances

Figure S4: The figure displays the VizRank widget. A maximum or exact number of features to be used when evaluating the data projections can be set, and the feature selection is performed by iteratively training and evaluating the performances of a k-Nearest Neighbor model, the parameters of which can be tuned in the "Settings" tab.



Attribute Evaluator

WrapperSubsetEval -B weka.classifiers.functions.PLSClassifier -F 5 -T 0.01 -R 1 -- -filter "weka.filters.supervised.attribute.PLSFilter -C 20 -M -A PLS1 -P center"

Search Method

BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) Class

weka.gui.GenericObjectEditor

weka.attributeSelection.WrapperSubsetEval

About

WrapperSubsetEval:

Evaluates attribute sets by using a learning scheme.

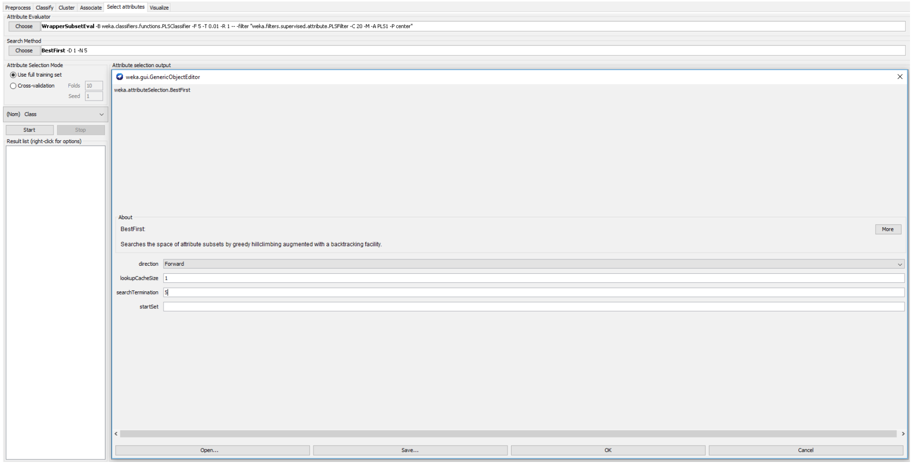
classifier **PLSClassifier** -filter "weka.filters.supervised.attribute.PLSFilter -C 20 -M -A PLS1 -P center"

folds 5

seed 1

threshold 0.01

Figure S5: The figure displays a typical workflow for feature selection in Weka. The feature selection algorithm can be chosen in the "Attribute evaluator" entry, while the parameters can be tuned by clicking onto the entry itself (the figure shows the window with the algorithm parameters). Moreover, the selection method can be set in the "Search Method" entry, along with some parameters (Figure S6). The best subset of features can be selected by testing the trained model onto the entire training dataset or via k-fold cross-validation.



Attribute Evaluator

WrapperSubsetEval -B weka.classifiers.functions.PLSClassifier -F 5 -T 0.01 -R 1 -- -filter "weka.filters.supervised.attribute.PLSFilter -C 20 -M -A PLS1 -P center"

Search Method

BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) Class

Attribute selection output

weka.gui.GenericObjectEditor

weka.attributeSelection.BestFirst

About

BestFirst:

Searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility.

direction: Forward

lookupCacheSize: 1

searchTermination: 5

startSet:

Figure S6: The figure displays the window that lists the parameters of the feature selection method, part of a typical workflow for feature selection in Weka.

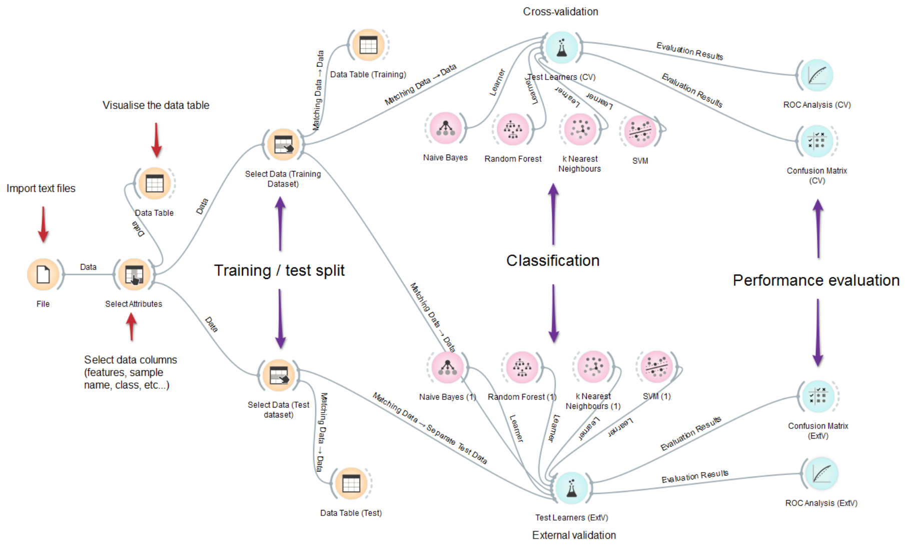
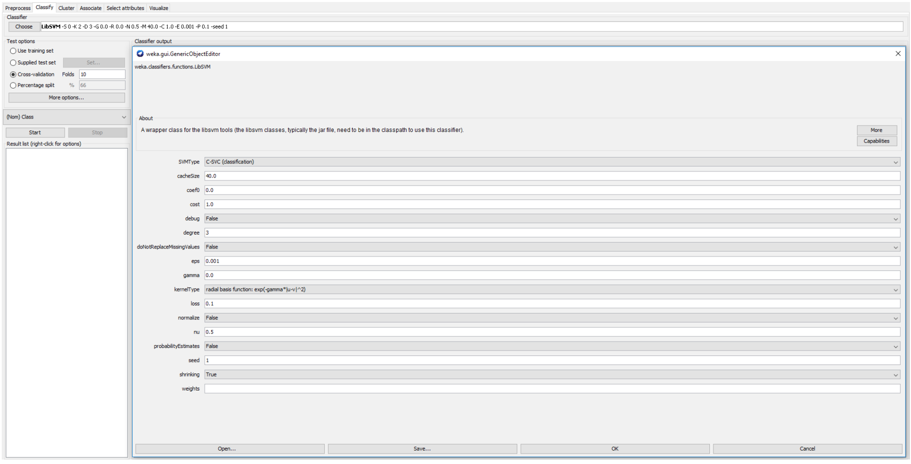


Figure S7: Typical workflow for classification analysis in Orange Canvas. After the data matrix has been imported, the software allows to select the data columns to be considered as attributes, class and meta-attributes, in order to include or exclude data columns in the analysis. The training/test split is performed by the “Select Data” operator, which conditionally filters the dataset, preserving only a subset of observations. The “Test Learners” operator takes this data (and possibly some other data as an external test set) and trains a classification model with it (Naive Bayes, Random Forest, k-NN and SVM in this instance), returning the classifier performances through the "ROC Analysis" and "Confusion Matrix" widgets.



Classifier

LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -seed 1

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Class

Result list (right-click for options)

Classifier output

weka.gui.GenericObjectEditor

weka.classifiers.functions.LibSVM

About

A wrapper class for the libsvm tools (the libsvm classes, typically the jar file, need to be in the classpath to use this classifier).

SVMType	C-SVC (classification)
cacheSize	40.0
coef0	0.0
cost	1.0
debug	False
degree	3
doNotReplaceMissingValues	False
eps	0.001
gamma	0.0
kernelType	radial basis function: exp(-gamma*|u-v|^2)
loss	0.1
normalize	False
nu	0.5
probabilityEstimates	False
seed	1
shrinking	True
weights	

Figure S8: Typical workflow for classification analysis in Weka. The classification algorithm can be chosen in the “Classifier” entry, while the parameters can be tuned by clicking onto the entry itself (the figure shows the window with the algorithm parameters, a Support Vector Machine in this instance). The classifier performances can be evaluated through the degree of concordance between the predicted class and the actual class, by testing the classifier onto the training dataset itself, by testing it onto an external test set, by performing a k-fold cross-validation or by performing a train/test split onto the dataset.

Chapter 3

A Support Vector Machine Classification of Thyroid Bioptic Specimens Using MALDI-MSI Data

Manuel Galli[1,*], Italo Zoppis[2,*], Gabriele De Sio[1], Clizia Chinello[1], Fabio Pagni[1], Fulvio Magni[1], and Giancarlo Mauri[2]

[1]Department of Medicine and Surgery, University of Milano-Bicocca, Monza Brianza, Italy ;

[2]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano, Italy

**These authors contributed equally to this work*

Advances in Bioinformatics

May 17 2016

doi: 10.1155/2016/3791214

ABSTRACT

Biomarkers able to characterize and predict multifactorial diseases are still one of the most important targets for all the “omics” investigations. In this context, Matrix-Assisted Laser Desorption/Ionization - Mass Spectrometry Imaging (MALDI-MSI) has gained considerable attention in recent years, but it also led to a huge amount of complex data to be elaborated and interpreted. For this reason, computational and *machine learning* procedures for biomarker discovery are important tools to consider, both to reduce data dimension and to provide predictive markers for specific diseases. For instance, the availability of protein and genetic markers to support thyroid lesion diagnoses would impact deeply on society due to the high presence of undetermined reports (THY3) that are generally treated as malignant patients. In this paper we show how an accurate classification of thyroid bioptic specimens can be obtained through the application of a state-of-the-art *machine learning* approach (i.e., *Support Vector Machines*) on MALDI-MSI data, together with a particular *wrapper feature selection algorithm* (i.e., *recursive feature elimination*). The model is able to provide an accurate discriminatory capability using only 20 out of 144 features, resulting in an increase of

the model performances, reliability, and computational efficiency. Finally, tissue areas rather than average proteomic profiles are classified, highlighting potential discriminating areas of clinical interest.

1. Introduction

Thyroid lesion diagnosis constitutes an important issue in terms of life quality of the affected patients. Currently, this pathology is diagnosed through cytomorphological evaluation of smears obtained after an ultrasound-guided fine needle aspiration biopsy (FNAB). A category of malignancy is then assigned to specimens according to the SIAPEC-IAP (Italian Society of Anatomic Pathology and Cytology) classification [1]. In particular, a category (ranging from THY1 to THY5 in the European system) is associated with the following lesions and groups of patients: inadequate withdrawal (THY1), benign lesions (THY2), lesions with unknown malignancy potential (THY3), and malignant lesions (THY4 and THY5). The general guidelines suggest that patients diagnosed as being of THY4 or THY5, along with the ones diagnosed as being of THY3, must undergo a total thyroidectomy and a consequent lifelong hormone replacing therapy, resulting in possible complications during or after surgery and in possible compliance issues during the patient's life. Surprisingly, 70% of the THY3 cases result benign after a deep histological evaluation after surgery [1], highlighting the diagnostic problem related to the undetermined reports (THY3).

The lack of protein and genetic biomarkers to reliably support thyroid lesion diagnoses led us to exploit the discriminative power of a *machine learning* technique (i.e., *Support Vector Machine*, SVM) applied to Matrix-Assisted Laser Desorption/Ionization-Mass Spectrometry Imaging (MALDI-MSI) data. MALDI-MSI data has already proven itself to be capable of highlighting differences in the proteomic profile of different types of thyroid lesions [2,3], further supporting our work. MALDI-MSI is an analytical technique that allows the study of the spatial distribution and relative abundance of a wide range of molecules directly on-tissue, without the need of any labeling or extraction processes that can possibly hinder both the molecular structure and the extraction yield of the analytes of interest [4]. For this reason MALDI-MSI has gained considerable attention in recent years and has been widely employed in several fields with successful results, from oncology and immunology to forensics and from pharmacology to the study of plants [5]. Although the advantages of MALDI-MSI are unquestionable for the explorative research, it also leads to file sizes of several gigabytes and more recently even terabytes of complex and high dimensional data from a single examined tissue slice. Computational analysis of MSI data and mining procedures are therefore challenging to be met [6].

Specifically, in this paper, we show how a *Support Vector Machine* based classification [7] can provide accurate discrimination of thyroid bioptic specimens using mass spectrometry imaging data, thus aiming at taking MALDI-MSI to the daily clinical practice to aid the clinical routine for diagnostic processes. Taking advantage of the general purpose applicability of the SVM model (broadly applied in both proteomics and more general biomolecular classification problems; see, e.g., [8] and [9], resp.) we provide accurate classification of THY3 patients to a benign or a malignant category. Moreover, to reduce the dimensionality of available data, we applied a *feature selection* algorithm (i.e., *recursive feature elimination*; see, e.g., [10]) to a derived dataset obtained through the generation of an average (representative) spectrum per patient.

The paper is laid out as follows. In Sections 2.1 and 2.2 we briefly describe the samples and the data acquisition process. In Section 2.4 we detail the preprocessing phase. In Section 3 we report the model construction and the “standard” classification process while in Section 4 we introduce the *pixel-by-pixel classification*, important to highlight potential discriminating areas of clinical interest. We show the results in

Section 6 and conclude, finally, in Section 7 by discussing our findings.

2. Materials and Methods

2.1. Patients

The study was conducted on leftover bioptic material collected at the Department of Pathology, University of Milano-Bicocca, Monza Brianza, Italy. A cohort of 43 subjects with the following characteristics (Table 1) was enrolled:

- 14 subjects diagnosed as being of THY2, 8 THY4, and 10 THY5 (for a total of 32 patients),
- 11 subjects diagnosed as being of THY3.

Patient number	Cytological diagnosis	Histological diagnosis
Patient 1	THY 2	Ben
Patient 2	THY 3	PTC
Patient 3	THY 4	PTC
Patient 4	THY 5	Ben
Patient 5	THY 2	PTC
Patient 6	THY 5	Ben

Patient number	**Cytological diagnosis**	**Histological diagnosis**
Patient 7	THY 2	Ben
Patient 8	THY 5	PTC
Patient 9	THY 3	PTC
Patient 10	THY 4	PTC
Patient 11	THY 2	Ben
Patient 12	THY 4	PTC
Patient 13	THY 3	Ben
Patient 14	THY 3	PTC
Patient 15	THY 4	PTC
Patient 16	THY 2	Ben
Patient 17	THY 2	Ben
Patient 18	THY 3	PTC
Patient 19	THY 2	Ben
Patient 20	THY 3	Ben
Patient 21	THY 4	PTC
Patient 22	THY 3	Ben
Patient 23	THY 5	PTC
Patient 24	THY 2	Ben
Patient 25	THY 4	PTC
Patient 26	THY 4	PTC
Patient 27	THY 2	Ben
Patient 28	THY 2	Ben
Patient 29	THY 2	Ben

Patient number	Cytological diagnosis	Histological diagnosis
Patient 30	THY 5	PTC
Patient 31	THY 5	PTC
Patient 32	THY 2	Ben
Patient 33	THY 5	PTC
Patient 34	THY 3	Ben
Patient 35	THY 2	Ben
Patient 36	THY 3	Ben
Patient 37	THY 3	Ben
Patient 38	THY 5	PTC
Patient 39	THY 5	PTC
Patient 40	THY 5	PTC
Patient 41	THY 3	Ben
Patient 42	THY 4	PTC
Patient 43	THY 2	Ben

Table 1: Table listing all the patients enrolled in the study, along with the cytological and histological diagnosis.
Ben: Benign lesions; PTC: Papillary Thyroid Carcinoma.

2.2. Acquisition of Mass Spectra

The cytological smears have been scanned through a ScanScope CS digital scanner (Aperio, Park Center Drive, Vista, CA, USA), to obtain a digitalized image of the specimen. After sample preparation, mass spectra were acquired using the ultrafleXtreme MALDI-TOF/TOF mass spectrometer (Bruker Daltonics GmbH, Bremen, Germany) in linear positive mode. All acquired spectra range from *m/z* 3000 to 25000, with a *raster* (namely, the spatial resolution) of 100 micrometers.

2.3. MALDI-MSI Data

Generally, a mass spectrometry imaging dataset consists of a “data cube” (Figure 1) resulting from the acquisition of one mass spectrum for each pixel of the digitalized tissue image. By considering a particular *mass-to-charge* (*m/z*) value, we can then represent the spatial distribution of the corresponding compound (with that specific *m/z*) by coloring each pixel according to its intensity values (i.e., relative abundance) at different spatial coordinates. In other words, for each *m/z* value in the spectrum, a molecular image showing the spatial distribution of the corresponding analyte is generated, possibly highlighting regions where the selected molecule localizes. Finally, spectra from specific regions of the sample can be exported and passed to the software for elaboration.

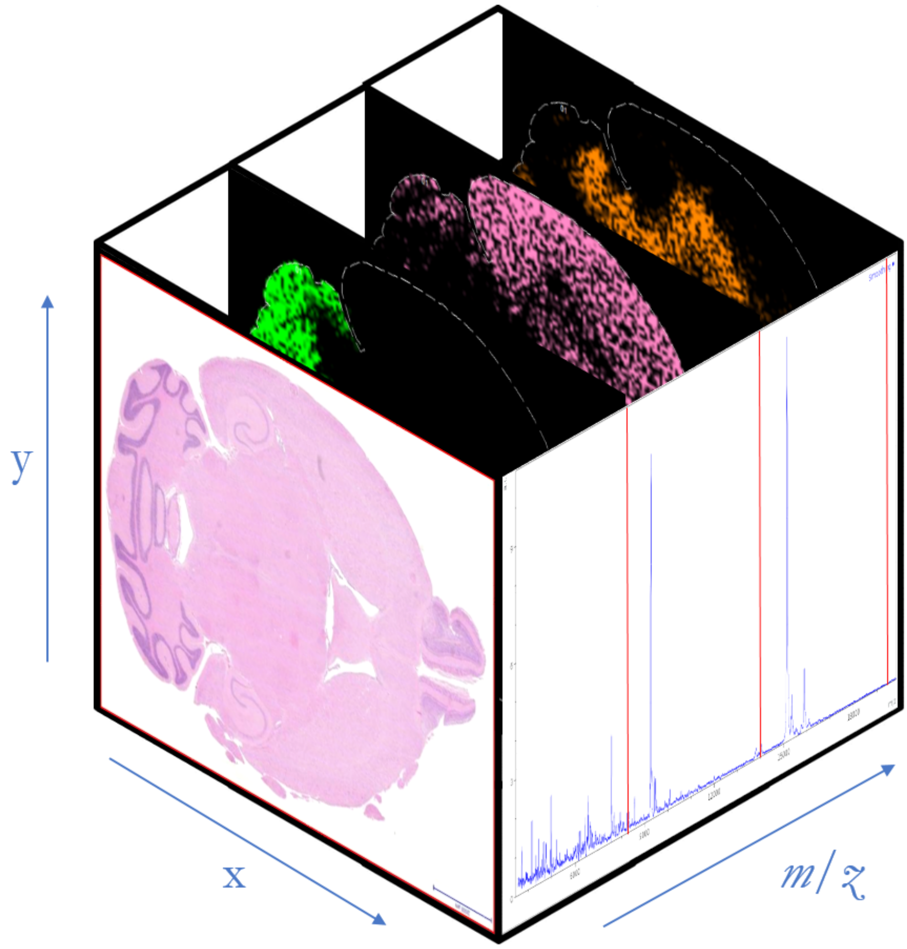


Figure 1: MALDI-MSI data cube. The intensity value of a specific analyte compound is localized as follows: x and y axes represent the spatial coordinates of the 2D digitalized tissue image (a mouse brain is shown in this example); the z-axis represents the mass-to-charge (m/z) ratio in the acquired spectra. 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.

2.4. Data Preprocessing

Raw data provided by MALDI instruments can be viewed also as a simple collection of independent spectra which are generally unaligned and noisy. Data preprocessing is a crucial step for allowing fair comparisons and reducing both technical and analytical variability or artifacts. To provide more reliable elaboration, we first applied the following steps.

(i) Baseline Subtraction and Smoothing. The baseline of a spectrum is a segment connecting points with the lowest intensities on which the entire spectrum lies. The baseline is essentially made of noise (electrical noise and chemical background generated by impurities), which, in turn, hinders the feature extraction process (*peak picking*). In this work the baseline subtraction process has been computed using the *TopHat* algorithm, while the denoising was performed using the *Savitzky-Golay* smoothing, in order to bring the spectra onto the x-axis and to present more defined peaks (thus allowing more reproducible *peak picking* selections [11]).

(ii) Normalization. Normalization is the process that consists in the multiplication of all the intensity values in the mass spectrum by a scaling factor, which results in an intensity axis broadening or narrowing. Here we applied the so-called *total*

ion count (TIC) method: all the intensities of each spectrum in the dataset are divided by the spectrum total current (i.e., the sum of all the intensities), providing each spectrum with the same integrated area under the curve [12].

(iii) Peak Picking. The *peak picking* was made using the *Median Absolute Deviation* (MAD) as a noise estimation method, with a *signal-to-noise* (*S/N*) ratio threshold of 3. The *peak picking* results in the selection of the highest *m/z*-intensity coordinates of the peaks in the spectra (i.e., features for the following selection phase) [13]. This leads to a massive reduction of the data dimensionality that will lead to a more computationally efficient analysis.

(iv) Peak Alignment and Filtering. All peaks have been aligned (with a tolerance of 2000 ppm) in order to prevent slightly analytical variations in the *m/z* values from being seen as distinct peaks. This ensures more consistent and coherent results, since possible artifacts in the identification of putative biomarkers are prevented from being generated. In addition, in order to remove false positive peaks coming from the noise, a filtering has been applied, resulting in keeping only the signals that are present in at least the 25% of all the spectra in the dataset.

The peak alignment and filtering have been performed on the entire dataset, as part of the preprocessing of the entire spectral data. Although this can potentially introduce some bias, especially in low-intensity peaks, closer to the noise and possibly not well resolved, this effect is compensated by the filtering, which was performed, for this reason, on the entire dataset. The *signal-to-noise* (*S/N*) ratio method (which we used) for *peak picking* is known to generate false positive peaks [6], and this is why the filtering is performed. Other peak picking methods, such as the orthogonal matching pursuit (OMP), which evaluates the shape of the peak rather than its intensity, are known to be more robust and reliable [6], but there are no R functions at the moment to perform peak picking with this algorithm. This could be an input for future work, to make peak picking more robust to peak shape and symmetry and to decrease the number of false positive peaks.

2.5 Peak-List Matrix and Data for Classification

The preprocessing step provided us with a *peak-list matrix*. This matrix with other elaborated data has been used to build and evaluate the inference model. We summarize the data we used as follows.

Dataset 1: Peak-List Matrix. As referred to above, this data is directly provided by the preprocessing step, yielding a number of aligned peaks of 144.

Average Profile Data. The obtained profiles were then used for the *average profile classification* as described in the next sections. In particular, for this task, the following two datasets were created.

Dataset 2: Training Set. It contains peak-list data from THY2, THY4, and THY5 patients.

Dataset 3: Validation (Test) Set. It contains peak-list data from THY3 patients.

3. Average Profile Classification

To obtain a classifier we applied sequentially the following steps.

(i) Recursive Feature Elimination. In this phase, we executed a wrapper feature selection process using the training set (dataset 2) as defined previously. To avoid overfitting and allow for the classifier to work properly, we applied a repeated (2 times) 10-fold cross-validation process with the recursive feature elimination (RFE) algorithm. In particular, to evaluate

the performances of the selected subsets of features, we iteratively applied a partial least squares (PLS) model (for implementation issue see R “caret” package [14]). In this way, we obtained a subset of 20 features, which, in turn, was submitted for further elaboration as described in the following step.

Feature selection decreases the risk of overfitting, especially with this reduced number of patients. When using individual spectra/pixels per patient, the risk of overfitting is reduced, but the algorithm can become slower and less efficient in terms of performances and classification capability (see the comparison of computational times in Table 2 and of classification performances in Tables 3 and 5; the process has been executed on a machine equipped with 16 GB of RAM, an Intel i7-4702mq CPU, and a 7200 rpm hard disk, on Ubuntu Linux): in fact, the mathematical formula that defines the model will be much more complicated. On the contrary, we want the algorithm to be fast and efficient, especially if an ensemble classifier is to be implemented in the future: when more algorithms are employed at the same time to vote for the class of the unknown sample, it is important that since the time taken by the process exponentially increases with the number of algorithms running, the classification is performed in reasonable time. This would also increase the translatability of the approach to the daily clinical routine. Finally, by retaining more peaks, the model can become more susceptible to variations in the peak intensity due to analytical variability and fluctuations in the instrument sensitivity and performances.

	Feature selection	**No feature selection**
RFE	75.656	//
SVM tuning and test	32.392	117.524

Table 2: Table displaying the difference in computational time taken by the classification process when employing the feature selection and when not. The tuning parameter grid is the same in both cases. Times are displayed in seconds and calculated by the R function system.time().

	Accuracy	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**ROC**
EV	0.273	0.000	1.000	0.000	0.273	0.500
2x 10-fold CV	0.567	0.000	1.000	0.000	0.567	0.500

Table 3: Validation performances of the SVM classifier without performing feature selection. In our case, the performances indicate the ability of the algorithm to correctly detect the benignity when the case is filed as THY3. EV: external validation; CV: cross-validation; PPV: Positive Predicted Value; NPV: Negative Predictive Value.

	Accuracy	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**ROC**
EV	0.818	0.750	1.000	1.000	0.600	0.875
2x 10-fold CV	0.713	0.625	0.775	0.740	0.767	0.778

Table 5: Validation performances of the SVM classifier after performing the RFE feature selection. In our case, the performances indicate the ability of the algorithm to correctly detect the benignity when the case is filed as THY3. EV: external validation; CV: cross-validation; PPV: Positive Predicted Value; NPV: Negative Predictive Value.

(ii) SVM Classification. A *Support Vector Machine* (SVM) model was trained using dataset 2 with the features provided by the *recursive feature elimination*. Moreover, the SVM was tuned to maximize the model capability, thus obtaining a classification with high performances. A 10-fold cross-validation was performed 2 times onto the training data to assess the reliability of the SVM. The trained classifier was then tested onto validation dataset 3 (THY3 patients), returning the classification performances based upon the degree of concordance between the predicted class and the actual class, in terms of sensitivity, specificity, positive predictive value (PPV), negative predicted value (NPV), and ROC AUC (area under the curve).

4. Pixel-by-Pixel Classification

After testing the SVM classifier onto the average proteomic profiles, we applied the trained model to predict the class of all the individual spectra in the MALDI-MSI dataset, which is one mass spectrum for each pixel: this results in a pixel-by-pixel classification, namely, the classification of tissue areas rather than the entire proteomic profile of a patient. Since for each spectrum the physical coordinates of the digitalized image are also retained, then it is also possible to color the corresponding pixels over the image. In other words, for each patient, a molecular image with pixels colored according to the class is shown, highlighting differently classified tissue areas.

In the classification of new (unknown) MSI data, the algorithm preprocesses the spectra in the same way as the training dataset and aligns the peaks from the new data to the ones used for building the model. The peak filtering is performed onto the unknown MSI data before running the pixel-by-pixel classification, not in the average profile classification, in order to discard the presence of false positive peaks picked by the MAD algorithm, when individual spectra/pixels per patient are used.

5. Implementation

All the conceptual procedures described in this paper have been coded using the R environment (https://www.r-project.org/). The spectra were formatted as imzML files [15], imported into R using the "*MALDIquantForeign*" package [13] and processed using the "*MALDIquant*" package [13].

6. Results

Our primary interest was to build an accurate model able to discriminate malignant from benign thyroid bioptic specimens. Our approach was empirical: we first designed a specific knowledge discovery process (Section 3) able to provide an accurate model for case versus control classification (i.e., THY2 versus THY4 and THY5). Then we evaluated the model performances onto a validation set (THY3) as described in the previous paragraphs.

Table 5 reports the performances obtained after a repeated (2 times) 10-fold cross-validation process (using dataset 2) onto the validation set containing only patients diagnosed as being of THY3 (dataset 3). The performances are based upon the degree of concordance between the predicted class and the actual class, in terms of sensitivity, specificity, positive predictive value (PPV), negative predicted value (NPV), and ROC (*Receiver Operating Characteristic*) AUC (area under the curve).

Specifically, Table 6 displays the difference between the class that was predicted by the model and the actual class provided by the histological analysis.

Sample	**Predicted class**	**True class**
Patient 2	Ben	Ben
Patient 9	PTC	PTC
Patient 13	Ben	Ben
Patient 14	PTC	PTC
Patient 18	PTC	PTC
Patient 20	PTC	Ben
Patient 22	Ben	Ben
Patient 34	Ben	Ben
Patient 36	Ben	Ben
Patient 37	Ben	Ben
Patient 41	PTC	Ben

Table 6: Discrepancy between the predicted class and the actual diagnosis.

A visualization of the obtained accuracy can also be given through the pie chart in Figure 2.

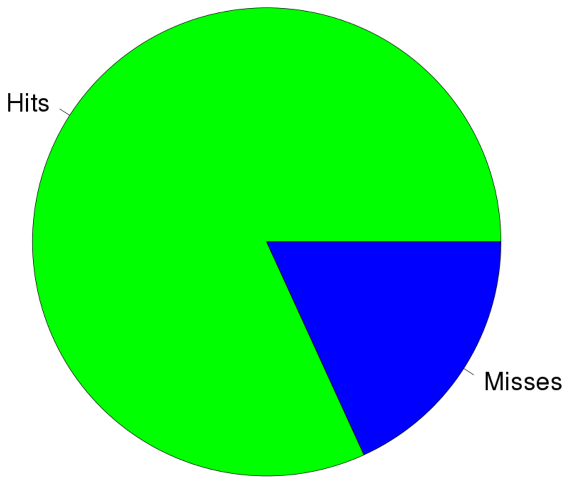


Figure 2: Graphical evaluation of the patient classification operated by the model. The green area is proportional to the amount of correctly classified patients, while the blue area corresponds to the amount of misclassifications.

The performances are further elucidated by the ROC curve, whose AUC (*area under the curve*) of 0.875 indicates a good capability of the model in assigning specimens to the correct class (Figure 3).

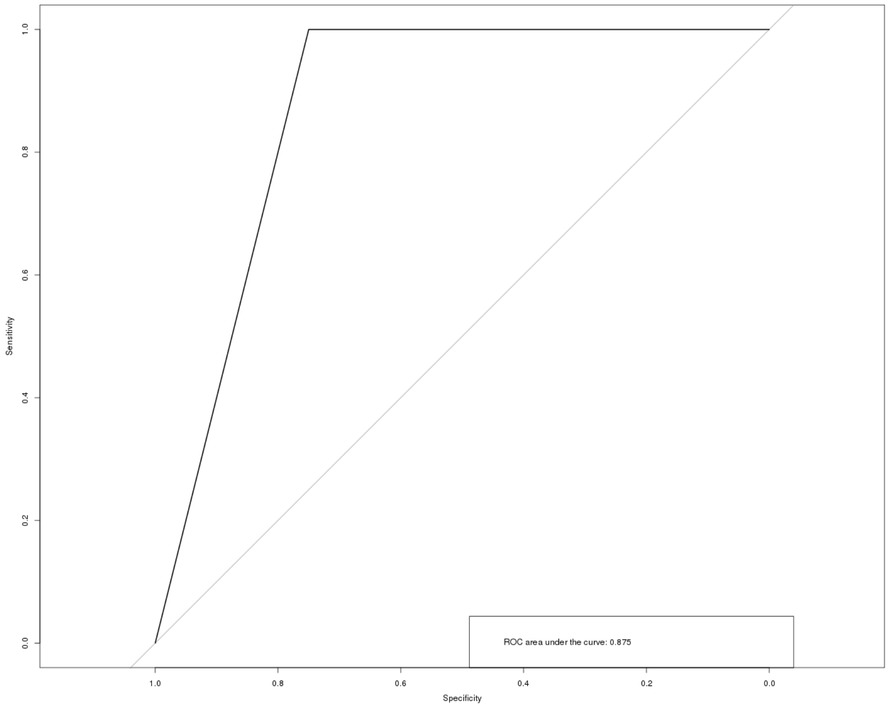


Figure 3: Receiver Operating Characteristic (ROC) curve computed by determining the number of true positive (sensitivity) and true negative (specificity) observations when employing the selected features.

Table 4 lists the parameters applied to the *Support Vector Machines* after the tuning (i.e., parameter optimization). The computational time of the automatic tuning process is clearly dependent on the range of values to be evaluated and the optimization method applied for the evaluation. In this case, we optimized the model parameters over a fixed set of default values (see, e.g., [14]) simply by taking the best resulting performance.

Feature selection	**Kernel**	**Cost**	**Epsilon**	**Gamma**
RFE	Radial	10	0.1	0.11
No RFE	Radial	10	0.1	1.11

Table 4: Tuning parameters of the support vector machines, with and without performing the feature selection. The best parameters are chosen according to the classification performance of the model.

As described above, MALDI-MSI data is represented by spectra corresponding to pixels of the digitalized tissue image. Instead of performing the classification onto the average proteomic profile only, this operation can be performed onto the individual spectra coming from the single patient as well. In this way, a *spectra-by-spectra* (corresponding to *pixel-by-pixel*) classification of the patient specimen can be obtained. Since spectra retain their spatial coordinates during the statistical analysis, it is also possible to color each pixel according to the inferred class (i.e., green for benign and red for malignant). This process resulted in the green and red area picture (Figure 4), providing a tissue area based classification rather than a standard profile classification of the entire proteomic profile.

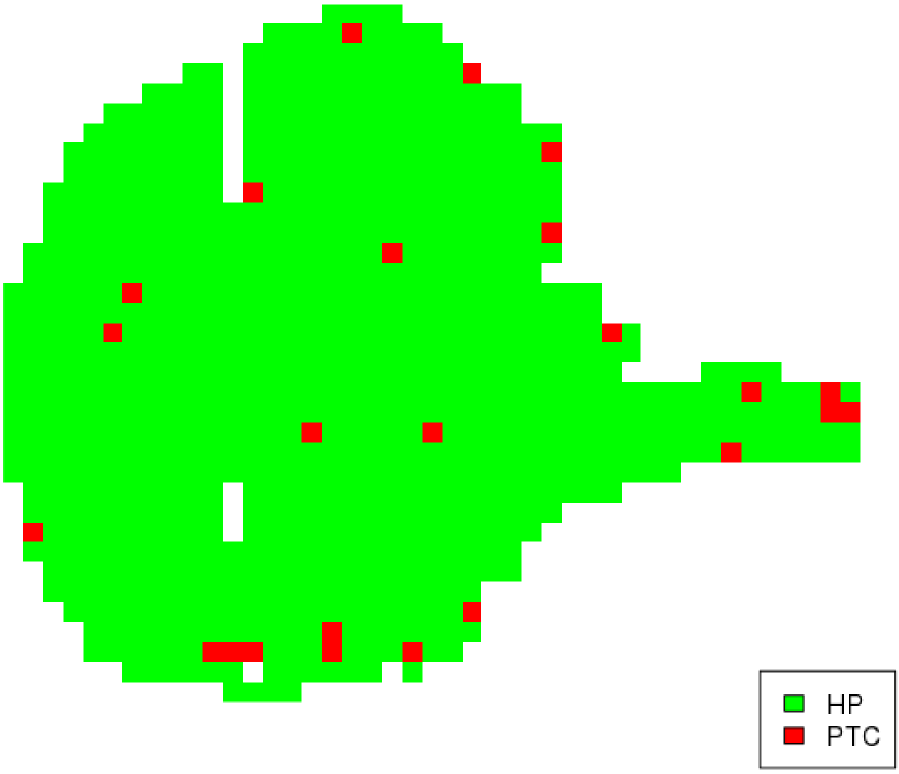


Figure 4: Pixel-by-pixel classification. An entire thyroid cytological smear is displayed. A mass spectrum was acquired for each pixel and the pixel-by-pixel classification has been applied. Green pixels correspond to spectra classified as benign (HP: hyperplastic), while red pixels correspond to malignant (PTC: papillary thyroid carcinoma) spectra.

7. Conclusion and Discussion

The work presented here shows the capability of MALDI-MSI to accurately classify unknown specimens obtained from the clinical routine. In this context, *machine learning* techniques (e.g., SVM) may be considered as a valuable approach able to exploit the full potentiality of the MALDI-MSI data, without the need of porting these findings to other clinical tests. This, in turn, allows MALDI-MSI to properly aid the diagnosis of specimens in the daily clinical practice. Importantly, given that MALDI-MSI looks at the sample at the molecular level, the possibility of performing a pixel-by-pixel classification constitutes a key point in the diagnostic process. In fact, areas highlighted by the inference model can represent regions that are undergoing molecular alterations that are not correlated with morphological changes or very tiny groups of cells that escaped the cytomorphological evaluation. Our results clearly suggest broader investigations either on different datasets or on different classification systems (i.e., ensemble classifiers). Moreover, the next studies will evaluate the possibility of MALDI-MSI to provide the information needed for identifying the correct subgroup of the pathology, to assess the disease progression, and to possibly detect the presence of the disease in the very early stages, providing concrete help in diagnoses. Finally, when more data is available, we will also exploit the possibility of classifying tissue specimens providing inference models directly trained on specific localized areas.

Acknowledgments

This study was supported by grants from the MIUR: FIRB 2007 (RBRN07BMCT_11), FAR 2010-2014, from iMODE-CKD (FP7-PEOPLE-2013-ITN), and in part by the COST Action (BM1104) Mass Spectrometry Imaging: New Tools for Healthcare Research.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

References

1. 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. doi: 10.1111/cyt.12085.

2. Mainini V., Pagni F., Garancini M., et al. An alternative approach in endocrine pathology research: MALDI-IMS in papillary thyroid carcinoma. Endocrine Pathology. 2013;24(4):250-253. doi: 10.1007/s12022-013-9273-8.

3. Pagni F., Mainini V., Garancini M., et al. Proteomics for the diagnosis of thyroid lesions: preliminary report. Cytopathology. 2015; 26(5):318-324. doi: 10.1111/cyt.12166.

4. Caprioli R. M. Imaging mass spectrometry: molecular microscopy for enabling a new age of discovery. Proteomics. 2014; 14(7-8):807-809. doi: 10.1002/pmic.201300571.

5. Aichler M., Walch A. MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. Laboratory Investigation. 2015;95(4):422-431. doi: 10.1038/labinvest.2014.156.

6. Trede D., Kobarg J. H., Oetjen J., Thiele H., Maass P., Alexandrov T. On the importance of mathematical methods for analysis of MALDI-imaging mass spectrometry data. Journal of Integrative Bioinformatics. 2012;9(1):p. 189.

7. Cristianini N., Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge, UK: Cambridge University Press; 2000.

8. Di Silvestre D., Zoppis I., Brambilla F., Bellettato V., Mauri G., Mauri P. Availability of MudPIT data for classification of biological samples. Journal of Clinical Bioinformatics. 2013;3(1, article 1) doi: 10.1186/2043-9113-3-1.

9. Cava C., Zoppis I., Gariboldi M., Castiglioni I., Mauri G., Antoniotti M. Combined analysis of chromosomal instabilities and gene expression for colon cancer progression inference. Journal of Clinical Bioinformatics. 2014;4(1, article 2) doi: 10.1186/2043-9113-4-2.

10. Guyon I., Gunn S., Nikravesh M., Zadeh L. Feature Extraction: Foundations and Applications. Vol. 207. Berlin, Germany: Springer; 2006.

11. Norris J. L., Cornett D. S., Mobley J. A., et al. Processing MALDI mass spectra to improve mass spectral direct tissue analysis. International Journal of Mass Spectrometry. 2007;260(2-3):212-221. doi: 10.1016/j.ijms.2006.10.005.

12. Deininger S.-O., Cornett D. S., Paape R., et al. Normalization in MALDI-TOF imaging datasets of proteins: practical considerations. Analytical and Bioanalytical Chemistry. 2011;401(1):167-181. doi: 10.1007/s00216-011-4929-z.

13. Gibb S., Strimmer K. Maldiquant: a versatile R package for the analysis of mass spectrometry data. Bioinformatics. 2012; 28(17):2270-2271. doi: 10.1093/bioinformatics/bts447.

14. Max Kuhn C. K., Wing J., Weston S., Williams A., Engelhardt A. Caret: Classification and Regression Training. 2012, https://cran.r-project.org/package=caret.

15. Schramm T., Hester A., Klinkert I., et al. ImzML—a common data format for the flexible exchange and processing of mass spectrometry imaging data. Journal of Proteomics. 2012;75(16):5106-5110. doi: 10.1016/j.jprot.2012.07.026.

Chapter 4

Combining multiple classifiers for the pixel-by-pixel classification of bioptic specimens by MALDI-MSI

Manuel Galli[1], Italo Zoppis[2], Giulia Capitoli[1], Stefania Galimberti[1], Giancarlo Mauri[2], and Fulvio Magni[1]

[1]Department of Medicine and Surgery, University of Milano-Bicocca, Monza Brianza, Italy ;

[2]Department of Computer Science, Systems and Communication, University of Milano-Bicocca, Milano, Italy

Bioinformatics

Submitted (November 2017)

ABSTRACT

Motivation: In the clinical routine, the diagnosis of diseases is performed via the cyto- or histo- morphological evaluation of bioptic specimens. Additionally, molecular diagnosis can be provided, when available, by immunohistochemistry, which, however, accounts for the presence of a limited set of biomarkers. Matrix-Assisted Laser Desorption/Ionization (MALDI) - Mass Spectrometry Imaging (MSI) is a powerful technology which allows the evaluation of the chemical composition of a tissue *in situ*, allowing for the detection of a large number of different molecules at the same time across the entire tissue section. The data consists of a data cube, composed of a *m/z* - intensity mass spectrum for each pixel of the digitalized tissue image. The aim of the study was to build a software pipeline for the exploitation of mass spectrometric data for clinical purposes, i.e the disease diagnosis of bioptic specimens at the molecular level.

Results: The software performs import and preprocessing of spectra, feature extraction, backwards feature selection (*Recursive Feature Elimination* - RFE), training and tuning of a series of classifiers, and finally classification of individual spectra, yielding a segmentation MS image (pixel-by-pixel

classification). The software has been tested onto clinical MSI data coming from the analysis of leftover bioptic specimens collected from the hospital. Spectra from Regions Of Interest (ROIs) drawn by the pathologist are imported, preprocessed and averaged to generate representative spectra. After the feature extraction phase (i.e. peak picking and peak alignment), a set of classifiers performs the RFE in order to select the best discriminatory features. Afterwards, each classifier is trained onto the reduced training set and fine-tuned in order to maximize the classification performances. Finally, each classifier predicts the outcome for each spectrum of a patient's data cube, and all the votes are combined together, according to the Bayesian framework, for determining the final outcome (pixel-by-pixel classification).

Availability: The software described in this paper is freely available as R script files (.R) and hosted at GitHub (at https://github.com/gmanuel89) under the GPLv3 license.

1 Introduction

1.1 The clinical issue

In the daily clinical practice, the diagnosis of diseases, in particular of cancer, is performed by pathologists via the morphological evaluation of bioptic material through cytology and histology [9,13,17]. Further investigation can be performed, when possible, by immunohistochemistry, which provides the possibility of a molecular diagnosis, through the detection of specific molecules with the aid of antibodies [14]. However, the implementation of multiple molecular biomarkers for the diagnostic process is still rather limited. Moreover, the lack of either consensus or guidelines regarding morphological boundaries and the lack of reliable biomarkers for molecular diagnosis leave cases unresolved [12].

Advanced technologies, such as genomics [1] and proteomics [11], constitute a remarkable assistance in the diagnostic procedure, providing molecular markers for a more objective and reliable diagnosis. Mass spectrometry, and in particular Mass Spectrometry Imaging (MSI), provides molecular insights that have been proven constituting a reliable assistant in the diagnostic process and of possible translatability to the daily clinical routine [5].

1.2 The MALDI-MSI technology

Matrix-Assisted Laser Desorption/Ionization (MALDI) - Mass Spectrometry Imaging (MSI) is the application of mass spectrometry which allows the pixel-by-pixel determination of the molecular composition of a tissue section and the evaluation of the localization of a wide spectrum of analytes, directly on tissue [2,8]. One of the biggest advantages of using mass spectrometry imaging is that local changes are not smoothed away or lost due to tissue homogenization but rather fully preserved in place. Therefore, more representative information can be extracted from tissue sub-areas, since small molecular changes are not lost when generating the average spectrum from that selected Region Of Interest (ROI) [2,8]. Additionally, the results yielded by MSI can be overlapped with histo-morphological features provided by staining the same tissue section used for the mass spectrometric analysis. This co-registration not only allows to extract spectra from specific ROIs defined by the pathologist for the definition of molecular signatures, but also provides a morphological and biological explanation to the results obtained from the MSI analysis.

The data collected from a MALDI-MSI analysis consists of a data cube, in which a mass spectrum (recording the presence and abundance of molecules through a *m/z* x *intensity* graph) is associated with each pixel of the digitalized tissue image.

1.3 Machine learning and MSI

The data obtained from MALDI-MSI has been found to be suitable for many applications in a great variety of fields, especially for clinical purposes [5]. Therefore, the application of state-of-the art machine learning algorithms onto MSI data seems a promising strategy for the exploitation of mass spectrometric information for the classification of clinical specimens.

However, the huge amount of data provided recent mass spectrometers and the intrinsic combinatorial nature of many computational problems require a careful and targeted application of both technological and methodological approaches. In this context, machine learning can provide useful methods to deal with such big and complex data. Moreover, distributed computational environments can be employed to shorten the time of computation and thus constitute a very efficient method to analyze such data.

2 System and Methods

2.1 Pixel-by-pixel classification

The aim of this work is to provide the user with a complete MALDI-MSI data analysis workflow, through an easy-to-use software series, that performs spectral import and preprocessing, feature selection, model training/tuning and finally pixel-by-pixel classification of MS datasets. The pixel-by-pixel classification is performed by an ensemble classification model, which can be trained and tuned before being applied to the MSI data.

The classification is performed on all the spectra of the MSI dataset (corresponding to pixels of the digitalized tissue image) collected from the analysis of an entire tissue section: each classifier of the ensemble returns the output of the classification, the final output being the result of a weighed vote among the classifiers. As already introduced, average molecular profiles are often not representative of the entire tissue section, since spectra coming from small and specific areas of tissue do not heavily contribute to the overall average spectrum and subtle molecular changes occurring in small areas of the biopsy are smoothed away by the averaging process. A pixel-by-pixel classification classifies tissue areas, by fully detecting molecular changes *in situ*, possibly bringing sub-areas of tissue to the attention of the pathologist. The generation of a segmentation image, by coloring pixels according to the predicted class of the corresponding spectrum, allows to identify putative areas of interest on a molecular basis, in a more reliable way, by employing the ensemble classification system.

2.2 Data size and parallel computation

The spectra collected from the mass spectrometer (ultrafleXtreme MALDI-TOF/TOF, Bruker Daltonics, Bremen, Germany), in reflectron positive mode, in a mass range of *m/z* 700-4000, with a spatial resolution of 100 µm (microns), are composed of around 60000 data points (*m/z* – intensity values) each. When imported into R, each spectrum is cached into the memory, allocating around 1 MB of space.

Given the amount of MS data, the computations cannot be performed with common local resources in reasonable time. For this study, a virtual machine hosted at the University of Milano-Bicocca was used, equipped with a 8-core CPU and 128 GB of RAM, in order to speed up the calculations by exploiting parallel computation.

3 Algorithm

The software employs the Bayesian framework in order to provide weights to the classifier votes, by determining the most probable outcome for each spectrum.

The Bayesian probability of a certain hypothesis given the data ($P(h|d)$) is proportional to the product of the prior probability ($P(h)$) and the likelihood probability ($P(d|h)$):

$$P(h|d) \propto P(d|h) \times P(h)$$

During the training phase, the classification performances (such as sensitivity, specificity, positive predicted value and negative predicted value) of each classifier are assessed during cross-validation and/or external validation (if an external dataset is provided) phase and computed for each class (i.e. levels of the response variable).

The *prior* probability is equal to the proportion of the observations in the training set with a selected outcome.

The *likelihood* is the probability of a certain hypothesis (i.e. the model predicting the real class) given the data (i.e. the class predicted by the model). Therefore, the likelihood probability is an indicator of the reliability of the classifier, by measuring the probability that the model will predict the correct class when applied to new unknown data.

The model classification performances are used as Bayesian probabilities, by consequently providing a weight to the classifier's vote. The Bayesian probabilities are calculated for each class, according to the conditional independence assumption. The sensitivity (or true positive rate – TPR, or recall) corresponds to the probability that the class predicted by the model is the actual class (P(d = 1 | h = 1) or P(d = 0 | h = 0)), while the false negative rate (FNR, equal to 1 - sensitivity) corresponds to the probability that the predicted class does not correspond to the actual class (P(d = 1 | h = 0) or P (d = 0 | h = 1)).

The final outcome is the class associated with the highest probability.

4 Implementation

4.1 Workflow

Import and preprocessing

All the spectra (in the data cube form) (formatted as imzML files [19] in the case of imaging data) are imported into R [16] using the “MALDIquantForeign” package [6]. In order to discard the analytical variability associated with the sample preparation and the electronic nature of the instrument, the spectra are subjected to preprocessing, comprising smoothing (algorithms: *Moving Average* or *Savitzky-Golay filter*), baseline subtraction (algorithms: *TopHat*, *SNIP*, *Convex Hull*, *Median*), normalisation (algorithms: *TIC*, *RMS*, *Median*) and alignment, by employing functions from the “MALDIquant” package [6]. The feature extraction phase, namely the peak picking (algorithms: *MAD* or *Super Smoother*) followed by the peak alignment, yields a n x p data matrix (peaklist matrix, number of observations: n, number of features: p) which, along with some additional clinical and demographical information, is submitted to the statistical analysis [5].

For this study, the spectra coming from homogeneous regions of each sample are averaged, by generating a representative average spectrum, which is used for feature extraction for the following phase.

Feature selection

Since overfitting issues can arise when training classifiers when *p* >> *n*, a feature selection step is performed in order to discard the redundant and invariant features and to keep only the most informative ones that have an actual impact on the classification. The *Recursive Feature Elimination* (RFE) (a backwards feature selection method) iteratively selects a subset of features and evaluates their impact onto the classification by fitting a model and giving weights to the features according to their relevance in the performances of the classification model itself. Finally, the most important features are preserved and carried through the statistical analysis: the feature selection algorithm can automatically select the optimal number of features according to the maximization of the model performances.

Each classifier of the ensemble (*Partial Least Squares*, *Support Vector Machines with Radial Basis Kernel Function*, *Support Vector Machines with Polynomial Kernel Function*, *Support Vector Machines with Linear Kernel Function*, *Naïve Bayes Classifier*, *k-Nearest Neighbor*, *Random Forest*) independently selects the best discriminatory features onto the same training set via the employment of the RFE algorithm: only the best performing features are preserved for the following tuning phase, in which the model is fine-tuned over a set of parameters (through a 3-time 10-fold cross-validation), in order to further maximize the classification capability. Each classifier is then cross-validated (3 times with a *k* of 10) in order to assess its performances and robustness, through the discrepancy between the predicted class and the actual class, by generating a confusion matrix and returning the performance parameters (such as sensitivity, specificity, positive predictive value and negative predictive value). All the statistical computations are performed by the employment of functions in the “caret” package [4].

Pixel-by-pixel classification

After every classifier has selected the best discriminatory features via the RFE and has been fine-tuned, all the classifiers are combined into an ensemble, in order to provide a more robust and reliable classification of specimens based onto MSI data. Each model individually predicts the outcome class of each spectrum of the data cube, and the final outcome is the result of a vote among the classifiers of the ensemble: the corresponding pixel is colored in red or in green according to the predicted class (red for malignant, green for benign).

4.2 Results

The software has been tested onto data coming from the MALDI-MSI analysis of bioptic specimens, provided as leftover material by the San Gerardo Hospital in Monza Brianza (Italy), in order to evaluate the potentiality of the mass spectrometric technology to provide reliable diagnosis at the molecular level, by confirming the molecular findings with the histological evaluation performed by the pathologist.

Dataset 1: Thyroid Tissue MicroArray (TMA)

Clinical relevance The diagnosis of thyroid malignancies is performed onto Fine Needle Aspiration Biopsies (FNABs), obtained from the cytological smear of bioptic material withdrawn from the patient's nodule, with the aid of ultrasound. The cyto-morphological evaluation of the smeared material determines the diagnosis of the patient. However, about 70% of the samples is filed as indeterminate for malignancy (THY3) and the patients undergo total thyroidectomy and lifelong hormone replacing therapy, as if they were affected by the malignancy [12]. Moreover, the inter-observer differences strongly affect the diagnostic process.

MALDI-MSI can assist the pathologists in the definition of a more objective molecular signature of malignancy and benignity, in order to make the diagnostic procedure more reproducible and reliable. In particular, Tissue MicroArrays (TMAs), by allowing the analysis of a high number of patients (between 80 and 100) at the same time, provide the throughput for strengthening the statistical analysis and, therefore, a better capability of detecting biomarkers to be employed for diagnosis.

Pixel-by-pixel classification by MALDI-MSI The software has been tested onto MSI data coming from the analysis of a formalin-fixed paraffin-embedded (FFPE) thyroid Tissue MicroArray (TMA). A TMA section is composed of multiple tissue cores, each chosen and taken from bioptic FFPE tissue by the pathologist.

The training set for the ensemble classification model comprised spectra coming from homogeneous TMA cores (6 benign and 6 malignant, affected by Papillary Thyroid Carcinoma - PTC). Around 500 mass spectra have been acquired from each core. Each spectrum underwent preprocessing, in order to enhance the biological information by discarding the analytical spectral variability: baseline subtraction with the *Statistics-sensitive Non-linear Iterative Peak-clipping* (SNIP) algorithm [18] (with 200 iterations) and normalization with the *Total Ion Count* (TIC) method. A representative average spectrum was generated for each tissue core, and peak picking was performed with the Friedman's *Super Smoother* algorithm [3] using a *signal-to-noise ratio* (S/N) threshold of 3 followed by peak deisotoping [15]. The feature selection (RFE) was performed by setting a number of features to retain of 30 and the accuracy as metric for selection of the best subset.

The ensemble pixel-by-pixel classification was then tested onto an entire TMA core, showing the presence of both benign and malignant tissue within the core (Figure 1). Histological evaluation performed by the pathologist has confirmed the molecular classification, proving the reliability of the molecular data for biological classification.

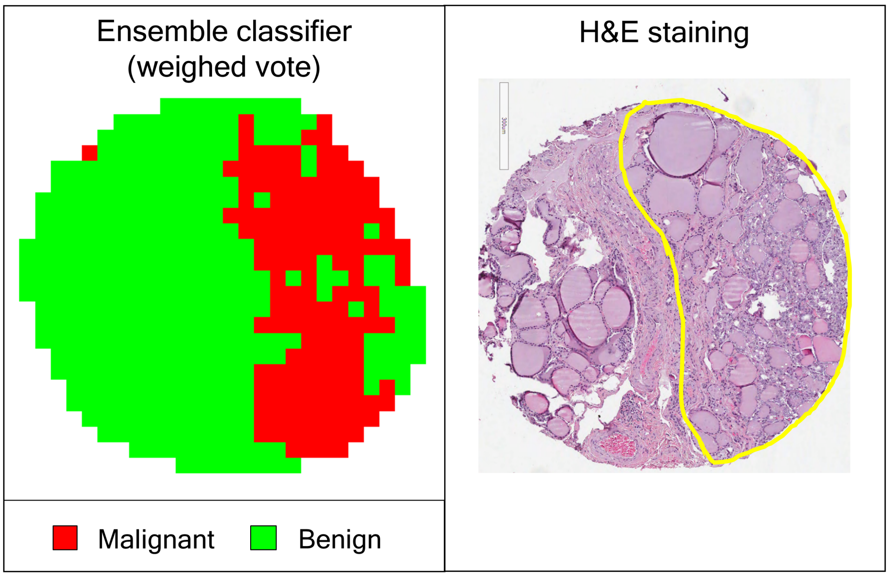


Figure 1. Comparison between the segmentation image computed by the software and the histology of the same tissue section. Red pixels correspond to spectra classified as malignant by the ensemble classification system, while green pixels belong to benign areas. On the histologically stained tissue section (Hematoxylin & Eosin staining) the area corresponding to the tumor is annotated in yellow by the pathologist.

	TPR	**TNR**	**PPV**	**NPV**	**Accuracy**
Pixel validation	1.000	0.651	0.705	1.000	0.810

Table 1. Validation performances of the ensemble classifier after performing the pixel-by-pixel classification. In our case, the performances indicate the ability of the algorithm to correctly detect the benignity, since it is clinically relevant to detect the benignity in the diagnostic phase. benignity in the diagnostic phase.

TPR: True Positive Rate (Sensitivity); TNR: True Negative Rate (Specificity); PPV: Positive Predicted Value; NPV: Negative Predictive Value.

Dataset 2: Medullary Thyroid Carcinoma (MTC)

Clinical relevance Medullary Thyroid Carcinoma (MTC) is a thyroid carcinoma originating from the parafollicular C cells, producing calcitonin. The majority of the cases (75%) is constituted by the sporadic form, the remaining percentage being represented by inherited MTC [10]. The diagnosis of MTC occurs after the cyto-morphological evaluation of a bioptic specimen: unfortunately, half of the malignancies are not diagnosed in the preoperative cytology phase [7]. The elevated levels of calcitonin produced by C cells in MTC cases can be used to aid the diagnostic procedure, even if no consensus has been reached onto the thresholds of the calcitonin levels to be employed to determine the MTC status [7]. Proteomics, and in particular MALDI-MSI, can reveal molecular alterations underlining the origin and development of MTC, to be used as diagnostic biomarkers.

Pixel-by-pixel classification by MALDI-MSI The software has been tested onto MSI data coming from the analysis of formalin-fixed paraffin-embedded (FFPE) thyroid tissue sections.

The pathologist annotated the Regions Of Interest (ROIs) on tissue, identifying the nodule and the healthy part of the tissue sections. Each region of interest was composed by around 3000 mass spectra. Each spectrum underwent preprocessing, in order to enhance the biological information by discarding the analytical spectral variability: baseline subtraction with the *Statistics-sensitive Non-linear Iterative Peak-clipping* (SNIP) algorithm [18] (with 200 iterations) and normalization with the *Total Ion Count* (TIC) method. A representative average spectrum was generated for each region of interest, and peak picking was performed with the Friedman's *Super Smoother* algorithm [3] algorithm using a *signal-to-noise ratio* (S/N) threshold of 3 followed by peak deisotoping [15]. Due to limitations in sample availability (5 tissue sections), two regions of interest, corresponding to one benign area and the tumor, were selected by the pathologist for each sample, yielding a total of two representative spectra for each patient. Therefore, the training set was composed of 5 benign and 5 malignant proteomic profiles, which were used to train and tune the classifiers. The feature selection (RFE) was performed by setting a number of features to retain of 30 and the accuracy as metric for selection of the best subset.

The classifier ensemble was then tested onto another sample, for the identification of the tumor area (Medullary Thyroid Carcinoma – MTC) within the tissue section (Figure 2).

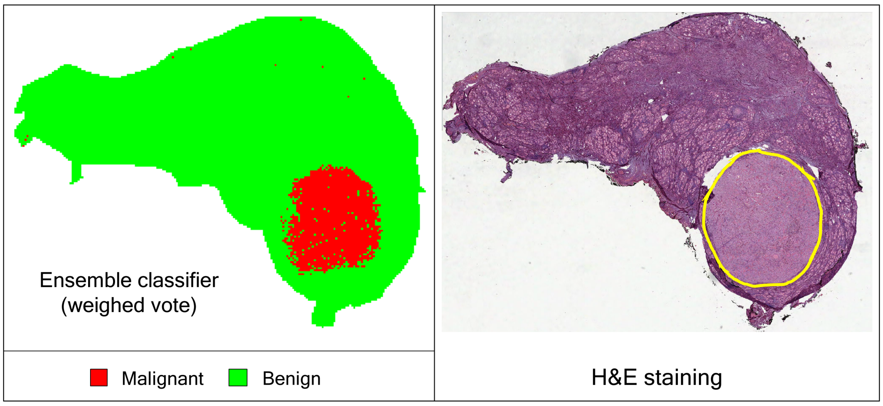


Figure 2. Comparison between the segmentation image computed by the software and the histology of the same tissue section. Red pixels correspond to spectra classified as malignant by the ensemble classification system, while green pixels belong to benign areas. On the histologically stained tissue section (Hematoxylin & Eosin staining) the area corresponding to the tumor is annotated in yellow by the pathologist.

	TPR	**TNR**	**PPV**	**NPV**	**Accuracy**
Pixel validation	0.746	0.999	0.997	0.947	0.954

Table 2. Validation performances of the ensemble classifier after performing the pixel-by-pixel classification. In our case, the performances indicate the ability of the algorithm to correctly detect the malignancy, since it is clinically relevant to detect the malignancy in the diagnostic phase. TPR: True Positive Rate (Sensitivity); TNR: True Negative Rate (Specificity); PPV: Positive Predicted Value; NPV: Negative Predictive Value.

5 Discussion and concluding remarks

In the daily clinical practice, the diagnosis of specimens is mostly performed by tissue staining followed by the cyto-morphological evaluation of cells, assigning a class to the sample according to features that the pathologists have established for diagnosis. The procedure also raises the question regarding the probability of inter-observer variability, which can lead to slightly different diagnosis performed in different centers by different pathologists.

Molecular diagnosis can be performed as well, for example with immunohistochemistry, by the employment of antibodies which bind to specific molecules that are differently expressed among different tissue classes. However, the procedure to release the specific antibody to the market for the routine use in the clinical practice can be rather long as it has to pass through a series of steps, and the employment of multiple antibodies for a single analysis is rather difficult and limited.

Matrix-Assisted Laser Desorption/Ionization (MALDI) - Mass Spectrometry Imaging (MSI) is a multivariate technique, which allows for the detection of a great variety of molecules on tissue: for each pixel of a digitalized tissue image, a mass spectrum is generated, recording the abundance of many different molecules simultaneously.

When different tissue sections from patients affected by a disease and control patients are analyzed, the pathologist can draw Regions of Interest (ROIs) from which to export mass spectra representative of that condition. The spectra undergo preprocessing before feature extraction, after which feature selection can be performed in order to discard the invariant and non-informative features from the dataset, which has a high number of features given the number of molecules detected by mass spectrometry. Therefore, classifiers trained onto mass spectrometric datasets can exploit the potentiality of the technique to make accurate predictions by using multiple features detected by the advanced instrumentations.

Despite the huge advantage of using machine learning in performing more objective diagnoses based upon molecular data, the training phase still relies upon the cyto-morphological and/or histopathological evaluation of expert pathologists for the correct assignment of the specimens to the corresponding disease. However, the high sensitivity and specificity of the mass spectrometric technology, coupled with a large patient cohort, provided also by the employment of formalin-fixed paraffin-embedded (FFPE) tissue specimens, allows to compensate the inter-observer variability in performing diagnoses by detecting subtle changes that characterize the disease and to train classifiers that can discriminate among different diseases based upon mass spectrometric features. Additionally, the employment of different classifiers at the same time, voting for patient classification, further strengthens the machine learning approach for the classification of clinical specimens.

The data collected from a MALDI-MSI analysis, in the form of a data cube, consists of thousands of spectra, which take around 1 MB of memory each when cached for processing. Common local resources are not enough to perform the proposed statistical analysis on the mass spectrometric dataset, which has to be handled by virtual machines, with higher computational power: the exploitation of parallel computing that relies on multiple cores provides a way of reducing the time of computation, therefore returning the output in reasonable time.

The software proposed in this work offers a complete workflow, from the import and preprocessing of spectra, to the training and tuning of classifiers, to the pixel-by-pixel classification of bioptic specimens. By exploiting the mass spectrometric data directly, the classification is performed at the molecular level, by accounting for the presence of several molecules simultaneously and by reducing the inter-observer variability in the cyto-morphological evaluation of stained specimens. By relying on parallel and distributed processing, the computation time can be shortened in such a way that the results are obtained in a relatively short time, given the amount of data that a mass spectrometry imaging analysis produces.

The application of state-of-the-art machine learning approaches on MALDI-MSI data seems to be a promising strategy for the molecular classification of bioptic specimens in the clinical practice.

The software has proven itself of being capable of predicting the presence of malignancy within a tissue section, by extracting the molecular signatures from a limited training set. When more patients are enrolled in the study, the classification system will become more robust and reliable, by reducing even further the risk of overfitting that arises from the employment of a small training set. Additionally, new methods for combining classifiers are under development, in order to tweak the weighing system in order to account more for a biologically relevant condition, e.g. when no false positives are allowed but some false negatives are. Finally, new methods for increasing the speed of computations are investigated, such as the exploitation of cloud computing clusters, in order to provide assistance to the pathologists by returning the classification output in reasonable time, even when dealing with such amount of complex data.

Acknowledgements

This study was supported by grants from the MIUR: FIRB 2007 (RBRN07BMCT_11), FAR 2010-2014; from iMODE-CKD (FP7-PEOPLE-2013-ITN-608332) and in part by the COST Action (BM1104) Mass Spectrometry Imaging: New Tools for Healthcare Research and by Fondazione Gigi & Pupa Ferrari Onlus.

References

1. Hila Benjamin, Temima Schnitzer-Perlman, Alexander Shtabsky, Christopher J. VandenBussche, Syed Z. Ali, Zdenek Kolar, Fabio Pagni, Dganit Bar, and Eti Meiri. Analytical validity of a microRNA-based assay for diagnosing indeterminate thyroid FNA smears from routinely prepared cytology slides. Cancer Cytopathology, 124(10):711–721, 2016.

2. Gile J. Caprioli RM, Farmer TB. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. Analytical Chemistry, 69(23):4751–4760, 1997.

3. Jerome H. Friedman. A Variable Span Smoother. Technical Report October, 1984.

4. Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. caret: Classification and Regression Training, 2017. R package version 6.0-76.

5. Manuel Galli, Italo Zoppis, Andrew Smith, Fulvio Magni, and Giancarlo Mauri. Machine learning approaches in MALDI-MSI: clinical applications. Expert Review of Proteomics, 13(7):685–696, 2016.

6. Sebastian Gibb and Korbinian Strimmer. MALDIquant: a versatile R package for the analysis of mass spectrometry data. Bioinformatics, 28(17):2270–2271, 2012.

7. Hassan M. Heshmati, Hossein Gharib, Jon A. Van Heerden, and Glen W. Sizemore. Advances and controversies in the diagnosis and management of medullary thyroid carcinoma, 1997.

8. Maciej Lalowski, Fulvio Magni, Veronica Mainini, Evanthia Monogioudi, Athanasios Gotsopoulos, Rabah Soliymani, Clizia Chinello, and Marc Baumann. Imaging mass spectrometry: a new tool for kidney disease investigations. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association, 28(7):1648–56, 2013.

9. Marco Manzoni, Gaia Roversi, Camillo Di Bella, Angela I. Pincelli, Vincenzo Cimino, Mario Perotti, Mattia Garancini, and Fabio Pagni. Solid cell nests of the thyroid gland: Morphological, immunohistochemical and genetic features. Histopathology, 68(6):866–874, 2016.

10. Masoumeh Mohammadi and Mehdi Hedayati. A Brief Review on The Molecular Basis of Medullary Thyroid Carcinoma. Cell Journal, 18(4):485–492, 2017.

11. F. Pagni, V. Mainini, M. Garancini, F. Bono, A. Vanzati, V. Giardini, M. Scardilli, P. Goffredo, A. J. Smith, M. Galli, G. De Sio, and F. Magni. Proteomics for the diagnosis of thyroid lesions: Preliminary report. Cytopathology, 26(5):318–324, 2015.

12. F. Pagni, M. Prada, P. Goffredo, G. Isimbaldi, S. Crippa, C. Di Bella, B. E. Leone, Maurizio Capra, Manuela Colombo, Rita Perego, Angela Ida Pincelli, Mario Perotti, Guido Grassi, Giovanni Colombo, Paolo Giannobi, Marcella Scardilli, and Vittorio Giardini. '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, 25(3):170–176, 2014.

13. Fabio Pagni, Marta Jaconi, Alberto Delitala, Mattia Garancini, Matteo Maternini, Francesca Bono, Alessandro Giani, and Andrew Smith. Incidental

Papillary Thyroid Carcinoma: Diagnostic Findings in a Series of 287 Carcinomas. Endocrine Pathology, 25(3):288–296, 2014.

14. Fabio Pagni, Marco Manzoni, Serena Buscone, and Biagio Eugenio Leone. ß-catenin as a morphoimmunohistochemical marker for the diagnosis of papillary thyroid carcinoma, 2015.

15. Kunsoo Park, Young Yoon Joo, Sunho Lee, Eunok Paek, Heejin Park, Hee Jung Jung, and Sang Won Lee. Isotopic peak intensity ratio based algorithm for determination of isotopic clusters and monoisotopic masses of polypeptides from high-resolution mass spectrometric data. Analytical Chemistry, 80(19):7294–7303, 2008.

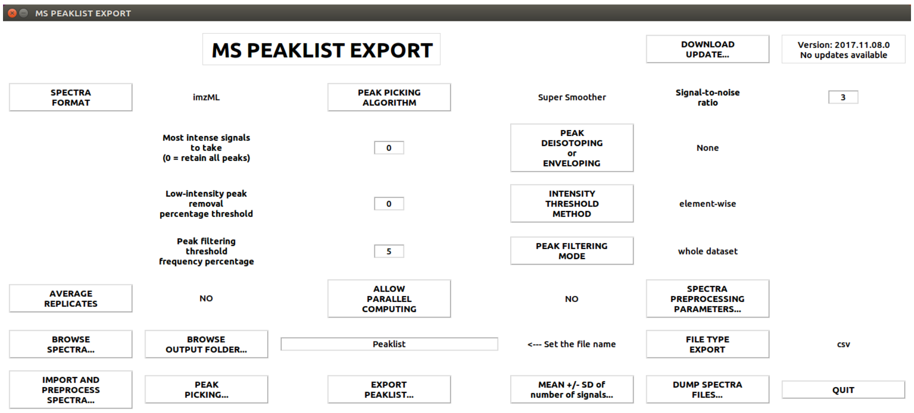
16. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2017.

17. Preetha Ramalingam. Morphologic, Immunophenotypic, and Molecular Features of Epithelial Ovarian Cancer. Oncology (Williston Park, N.Y.), 30(2):1–15, 2016.

18. C. G. Ryan, E. Clayton, W. L. Griffin, S. H. Sie, and D. R. Cousens. SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. Nuclear Inst. And Methods in Physics Research, B, 34(3):396–402, 1988.

19. Thorsten Schramm, Alfons Hester, Ivo Klinkert, Jean Pierre Both, Ron M A Heeren, Alain Brunelle, Olivier Laprévote, Nicolas Desbenoit, Marie France Robbe, Markus Stoeckli, Bernhard Spengler, and Andreas Römpp. ImzML - A common data format for the flexible exchange and processing of mass spectrometry imaging data. Journal of Proteomics, 75(16):5106–5110, 2012.

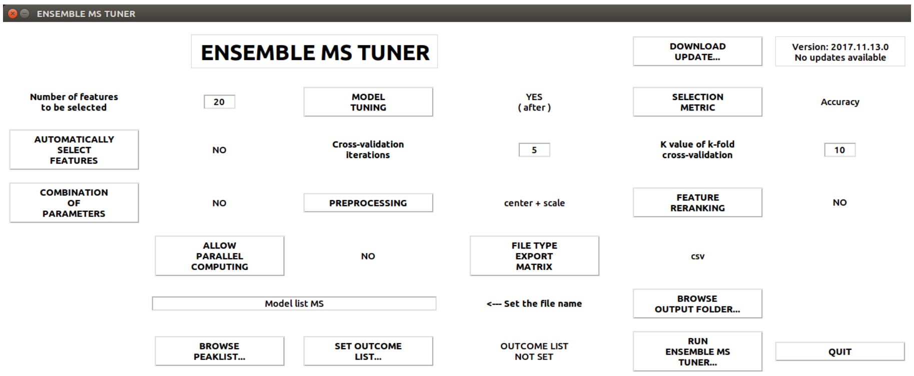
Supplementary material



MS PEAKLIST EXPORT

	MS PEAKLIST EXPORT			DOWNLOAD UPDATE...	Version: 2017.11.08.0 No updates available
SPECTRA FORMAT	imzML	PEAK PICKING ALGORITHM	Super Smoother	Signal-to-noise ratio	3
	Most intense signals to take (0 = retain all peaks)	0	PEAK DEISOTOPING or ENVELOPING	None	
	Low-intensity peak removal percentage threshold	0	INTENSITY THRESHOLD METHOD	element-wise	
	Peak filtering threshold frequency percentage	5	PEAK FILTERING MODE	whole dataset	
AVERAGE REPLICATES	NO	ALLOW PARALLEL COMPUTING	NO	SPECTRA PREPROCESSING PARAMETERS...	
BROWSE SPECTRA...	BROWSE OUTPUT FOLDER...	Peaklist	<--- Set the file name	FILE TYPE EXPORT	csv
IMPORT AND PREPROCESS SPECTRA...	PEAK PICKING...	EXPORT PEAKLIST...	MEAN +/- SD of number of signals...	DUMP SPECTRA FILES...	QUIT

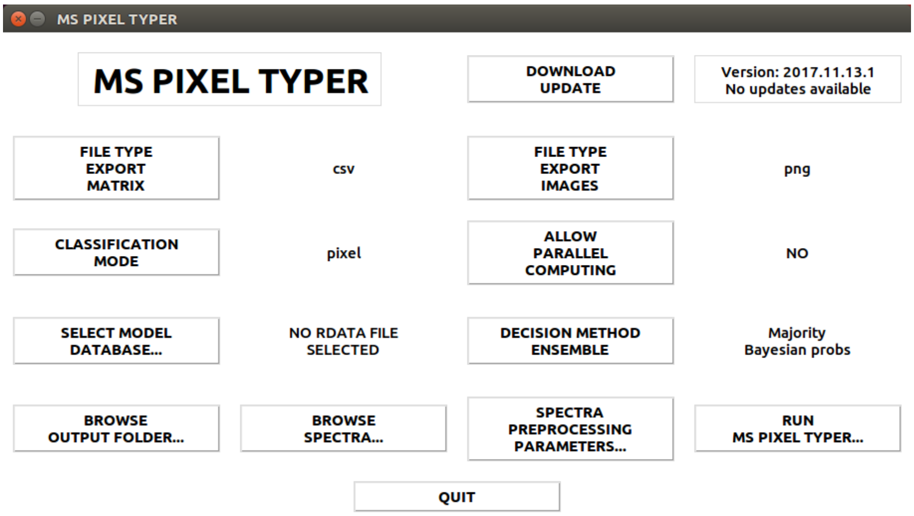
Figure S1: MS PEAKLIST EXPORT: generates and exports the statistical peaklist matrix from spectral files. The graphical user interface is generated through the Tcl/Tk package.



ENSEMBLE MS TUNER

	ENSEMBLE MS TUNER			DOWNLOAD UPDATE...	Version: 2017.11.13.0 No updates available
Number of features to be selected	20	MODEL TUNING	YES (after)	SELECTION METRIC	Accuracy
AUTOMATICALLY SELECT FEATURES	NO	Cross-validation iterations	5	K value of k-fold cross-validation	10
COMBINATION OF PARAMETERS	NO	PREPROCESSING	center + scale	FEATURE RERANKING	NO
	ALLOW PARALLEL COMPUTING	NO	FILE TYPE EXPORT MATRIX	csv	
	Model list MS		<--- Set the file name	BROWSE OUTPUT FOLDER...	
	BROWSE PEAKLIST...	SET OUTCOME LIST...	OUTCOME LIST NOT SET	RUN ENSEMBLE MS TUNER...	QUIT

Figure S2: ENSEMBLE MS TUNER: trains and tunes a series of classifiers onto the peaklist matrix features. The graphical user interface is generated through the Tcl/Tk package.



Version: 2017.11.13.1
No updates available

FILE TYPE EXPORT MATRIX csv

FILE TYPE EXPORT IMAGES png

CLASSIFICATION MODE pixel

ALLOW PARALLEL COMPUTING NO

SELECT MODEL DATABASE... NO RDATA FILE SELECTED

DECISION METHOD ENSEMBLE Majority
Bayesian probs

Figure S3: MS PIXEL TYPER: performs the pixel-by-pixel classification of MS images by generating a MS segmentation image (green/red pixels) resembling the spectral classification. The graphical user interface is generated through the Tcl/Tk package.

Chapter 5

Generating iMatrixSpray methods for MALDI-MSI analyses

Manuel Galli[1,*], Andrew Smith[1,*], Isabella Piga[1,*], Martina Stella[1,*], and Fulvio Magni[1]

[1]Department of Medicine and Surgery, University of Milano-Bicocca, Monza Brianza, Italy

** These authors contributed equally to this work*

This work is ongoing and will be submitted as a technical note to the *Analytical Chemistry* journal.

ABSTRACT

Matrix-Assisted Laser Desorption/Ionization (MALDI) – Mass Spectrometry Imaging (MSI) is an advanced technology capable of detecting small molecular changes occurring in tissue sections while preserving the spatial localization of such modifications. In order to provide this powerful data, to be employed for diverse clinical purposes, the sample preparation procedure is very delicate and implies the deposition of an organic compound, called the matrix, in order to allow for analyte extraction and detection by the mass spectrometric instrumentation. The deposition of the matrix must occur throughout the tissue in extremely fine layers, to achieve both the extraction of the molecules and the preservation of the localization of such molecules by generating small droplets of matrix solution which become small crystals on-tissue. The iMatrixSpray, being a robotic device composed of an arm moving onto a surface, aims at achieving a balance between the generation of fine layers of small matrix crystals and analyte extraction. However, it does not come with a software with an easy-to-use graphical user interface (GUI) for the extensive definition of the spraying parameters, in order to allow the user to set each individual parameter of the spraying

method depending on the purpose. The knowledge of the G-code language is required in order to achieve the specific aims of each laboratory. The work described in this technical note provides the scientific community with an easy-to-use software that allows to automatically generate G-code files, serving as methods for the iMatrixSpray device, to be employed by laboratories utilizing the spraying device, without any knowledge of the G-code language. Finally, the technical note provides the scientific community with example methods, generating by using the software, to be used as starting point by laboratories using the iMatrixSpray device.

1. Introduction

MALDI Mass Spectrometry

Matrix-Assisted Laser Desorption/Ionization (MALDI) Mass Spectrometry (MS) allows to determine the composition of a sample via the ionization of a wide spectrum of analyte molecules, aided by the energy provided by a laser beam. Due to the incompatibility between the laser wavelength and the absorbance wavelength of molecules, the sample must be dissolved in a compound, called the matrix, which, after solvent evaporation, co-crystallizes with the analyte molecules, absorbs the energy from the laser and transfers it to the analytes, promoting ionization. The ionization process takes place through proton transfer, and, according to the type of molecules to be detected, different matrix compounds can be employed: acidic matrices provide protons to the analytes, while basic matrices absorb protons from analytes, generating cations and anions respectively. Since MALDI is a soft ionization technique, no fragmentation occurs, and both intact proteins and endogenous peptides can be detected, as well as lipids, metabolites and drugs.

Matrix deposition: profiling vs imaging

When employing tissue sections, two MALDI-MS approaches can be applied. Profiling (MALDI-MSP) aims at obtaining a representative profile from the entire tissue section, or from sub-regions: intra-section or intra-region molecular differences are smoothed away in the averaging process and possibly discarded during the statistical analysis. In Imaging (MALDI-MSI), on the other hand, one mass spectrum for each pixel of the digitalized tissue image is acquired: in this way, local molecular changes are fully preserved in place and taken into account during the statistical analysis. From an imaging acquisition, a representative profile can be obtained by averaging spectra corresponding to specific pixels of the tissue image. Depending on the purpose to be achieved, the matrix deposition is performed accordingly. Profiling aims at maximizing analyte extraction, through tissue homogenization or by the deposition of large droplets of matrix solution on tissue. In this way, however, local changes are lost due to molecules dissolving into the matrix solution and diffusing within the droplet. Imaging, on the other hand, aims at performing an *in situ* extraction, by depositing small droplets of matrix solution on tissue, through spraying mostly, resulting in the full preservation of molecular changes in place.

2. Materials and Methods

The iMatrixSpray device

The iMatrixSpray device (https://imatrixspray.com/) is a robotic system (Delta Robot) which uses a moving arm to spray a solution onto a defined area. The plate onto which the samples are placed can be equipped with a heat bed, in order to control the temperature at which the sample is and at which the on-tissue co-crystallization process takes place. The spray arm is equipped with a capillary, connected to different vials through a multi-channel valve: one vial is reserved to waste coming from the rinsing of the components of the device, one vial is reserved to the rinsing solution itself and three vials are dedicated to the solutions to be sprayed. Therefore, up to three different solutions can be sprayed onto the sample within the same run.

Method generation for the iMatrixSpray device

The iMatrixSpray device provides a simple web application to allow the user to edit the parameters of spraying, such as the height of the needle, the speed of movement of the arm, the solution to be used, the density of the solution on tissue (directly related with the strength of the spray), the number of spraying cycles and the distance between the lines of spray. In order to further tweak the spraying method (e.g. reducing the number of rinsing cycles, adding pauses between the sprays, extending the drying time, changing the direction of spraying, changing the area of spraying, etc...), a proper G-code file must be generated, including all the commands that serve as instructions for the device to move the arm properly and spraying the correct amount of material on the area of interest.

3. Results

The in-house developed iMatrixSpray Method G-code Generator

In order to provide the end user with the capability of generating G-code files to be used as spraying methods for the iMatrixSpray device, allowing also for an extensive tweak of such methods, an easy-to-use application has been developed, written in Python (v3.6), that guides the end user in the process through a simple and intuitive graphical user interface. The software is hosted on GitHub, at https://github.com/gmanuel89/iMatrixSpray, under the GPLv3 license: software updates are deployed through the GitHub platform.

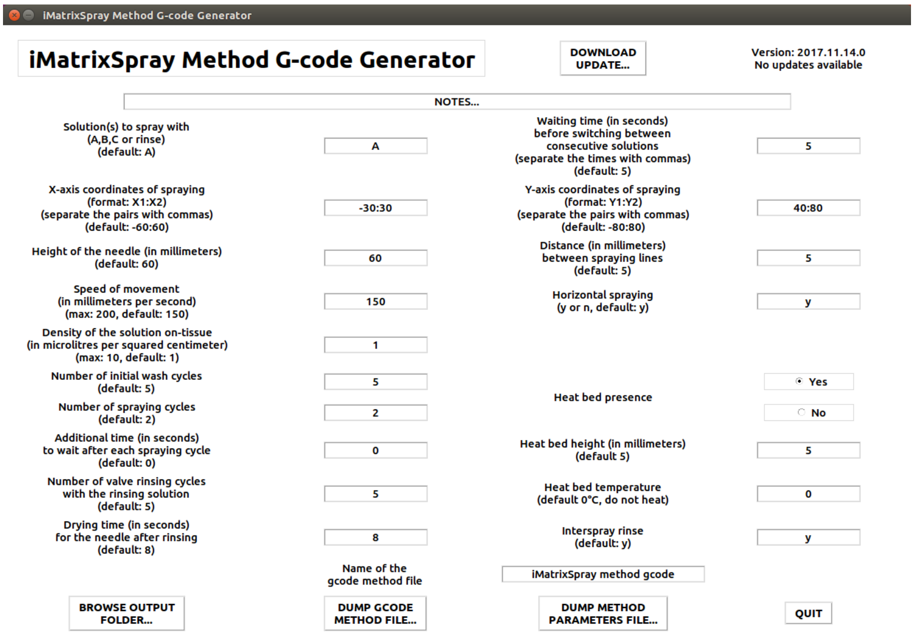
Through a simple graphical user interface, the user can:

- Define the solution to use for spraying. If more than one solution is specified, one full method is generated for each solution and all the methods are bundled together in the same G-code file, in such a way that one solution is sprayed after the other in the specified order with the associated method.
- If more than one solution is specified, the waiting time between two consecutive spraying methods can be set.
- The X and Y coordinates of the vertices of the spraying area can be specified, in order to limit the amount of solution to be used to only the area where the sample is placed.
- The Z coordinate is automatically determined according to the value specified for the height of the needle. The calculations account also for the possible presence of the heat bed.
- The distance between the lines of spraying defines the coverage of the solution on tissue.

- The speed of movement indicates how fast the arm is moving during the spray.
- The direction of spraying can be set to be either horizontal (along the X-axis) or vertical (along the Y-axis).
- The on-tissue density defines the strength of the spray, i.e. the amount of solution to be deposited for each line of spray.
- The number of spraying cycles defines how many times the tissue must be entirely covered by the solution. An additional time after each spraying cycle can be set, in order to leave the machine still before starting to spray again.
- The presence of a heat bed can be specified, along with the thickness of the heat bed itself, in order to adjust the height of the needle (Z coordinate) accordingly. Moreover, the temperature for the heat bed can be set.
- Additional tweaks can be applied to the method, such as changing the number of valve rinsing cycles, the number of initial wash cycles and the drying time. This can be useful in order to shorten the time of spraying,

especially when different methods with the same solution are generated.

When all the parameters are set, the method can be saved as a G-code file (.gcode) to be used by the iMatrixSpray device as method. Additionally, a CSV file with the method parameters can be generated along with the method file.


iMatrixSpray Method G-code Generator
iMatrixSpray Method G-code Generator
DOWNLOAD UPDATE...
NOTES...
BROWSE OUTPUT FOLDER...
DUMP GCODE METHOD FILE...
DUMP METHOD PARAMETERS FILE...
QUIT

Version: 2017.11.14.0
No updates available

Solution(s) to spray with (A,B,C or rinse) (default: A)	A	Waiting time (in seconds) before switching between consecutive solutions (separate the times with commas) (default: 5)	5
X-axis coordinates of spraying (format: X1:X2) (separate the pairs with commas) (default: -60:60)	-30:30	Y-axis coordinates of spraying (format: Y1:Y2) (separate the pairs with commas) (default: -80:80)	40:80
Height of the needle (in millimeters) (default: 60)	60	Distance (in millimeters) between spraying lines (default: 5)	5
Speed of movement (in millimeters per second) (max: 200, default: 150)	150	Horizontal spraying (y or n, default: y)	y
Density of the solution on-tissue (in microlitres per squared centimeter) (max: 10, default: 1)	1		
Number of initial wash cycles (default: 5)	5	Heat bed presence	● Yes
Number of spraying cycles (default: 2)	2		○ No
Additional time (in seconds) to wait after each spraying cycle (default: 0)	0	Heat bed height (in millimeters) (default 5)	5
Number of valve rinsing cycles with the rinsing solution (default: 5)	5	Heat bed temperature (default 0°C, do not heat)	0
Drying time (in seconds) for the needle after rinsing (default: 8)	8	Interspray rinse (default: y)	y
	Name of the gcode method file	iMatrixSpray method gcode	

Figure 1: The graphical user interface (generated through the Tkinter package) of the iMatrixSpray Method G-code Generator application.

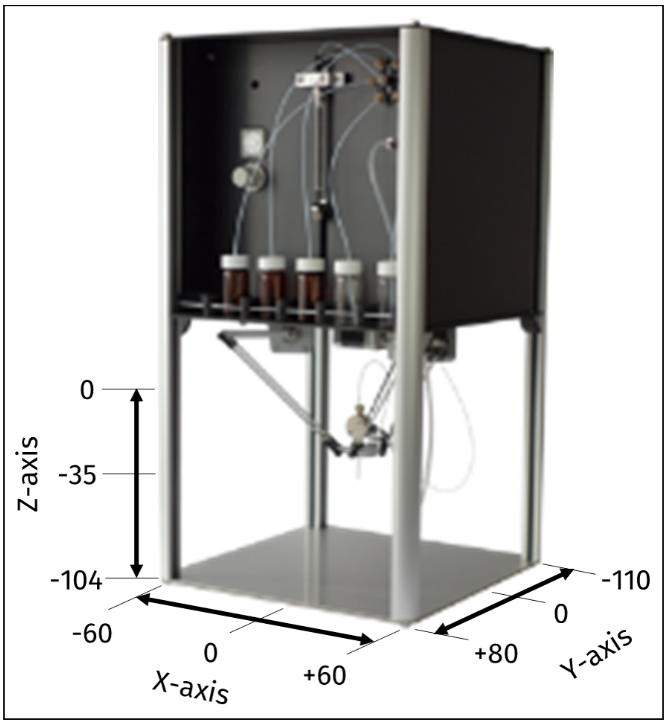


Figure 2: Graphical representation of the iMatrixSpray coordinates of spraying.

4. Concluding remarks

Recent advancements in mass spectrometry, involving mainly the development of more powerful and sophisticated instruments, require adequate sample preparation methods, especially trypsin and matrix deposition, to be employed, in order to enhance such instrumental evolution. An inadequate deposition of matrix, for example, will result in either a poor analyte extraction or a spatial delocalization of extracted molecules. When it comes to trypsin, on the other hand, digestion might not occur efficiently throughout the tissue section, resulting in a limited and unreproducible release of peptides. The problems encountered during sample preparation reflect onto the obtained data, which yields incorrect or misleading results, even when data analysis is performed correctly.

The iMatrixSray device has been proposed for solving the issue of matrix and trypsin deposition, by depositing fine layers of solution on tissue, guaranteeing a fair compromise between analyte extraction and preservation of localization. However, the iMatrixSpray device does not come with a full-featured application that can allow the end user to tweak the spraying method parameters extensively, at the moment. Therefore, the work proposed in this technical note provides the scientific community with the possibility to easily generate G-code methods and with example methods to be used as a starting point by laboratories which aim at employing the iMatrixSpray device for sample preparation for mass spectrometry imaging analyses.

Chapter 6

Summary, conclusions and future perspectives

6.1 Summary

Introduction/Rationale: Matrix-Assisted Laser Desorption/Ionization (MALDI) - Mass Spectrometry Imaging (MSI) is a powerful technology which enables the molecular composition of a specimen to be evaluated directly *in situ*. The acquired data has been found to be particularly suitable for clinical purposes and it is hoped that MALDI-MSI can eventually be used to provide diagnostic assistance in particularly difficult cases. The aim of this work is to provide the user with a complete MALDI-MSI data analysis workflow, through an easy-to-use software interface. The end-user should have the capability to perform the import and preprocessing of spectra, feature extraction and selection, model training/tuning and, finally, pixel-by-pixel classification of MS images, with the final output being the molecular classification of bioptic specimens.

Methods: The software program makes use of an ensemble of classifiers, rather than a single classification algorithm, to classify patients. Each classifier independently selects the most discriminatory features by discarding invariant and redundant features through the application of the *Recursive Feature Elimination* (RFE) algorithm. These undergo cross-validation and tuning over a set of parameters, to further assess and maximize the performances. Finally, the classification is performed on all the spectra collected from the analysis of the entire biopsy, with each algorithm contributing to the final report with a vote (weighed according to the Bayesian framework): pixels corresponding to spectra are coloured according to the predicted class, by generating a red/green MS segmentation image that resembles the classification.

Results: The software has been tested using data acquired following the MALDI-MSI analysis of formalin-fixed paraffin-embedded (FFPE) bioptic specimens. The molecular diagnosis was then correlated and confirmed following histological evaluation performed by a pathologist. The software was able to successfully detect benign and malignant tissue cores within a thyroid Tissue MicroArray (TMA). In order to extract the molecular signature of benignity and malignancy to be applied for classification, the training set was composed of histologically homogeneous tissue cores. The application of this classification highlighted the presence of a heterogeneous core (affected by Papillary Thyroid Carcinoma – PTC). From a molecular standpoint the core was determined to contain both benign and malignant regions, whilst initial histological evaluation indicated the presence of only benign cells. This tissue heterogeneity was then confirmed by further histological evaluation. The software has also been applied for the molecular detection of Medullary Thyroid Carcinoma (MTC). The training set was composed of benign and malignant tissue sub-areas, which were highlighted by the pathologist as being homogeneous in terms of cell composition and type. When tested on clinical MTC tissue specimens, the software was able to successfully highlight the area corresponding to the tumour nodule.

Conclusions/Novelty: MALDI-MSI, coupled with the application of state-of-the-art *machine learning* algorithms, can potentially provide assistance during the diagnostic process by evaluating the molecular alterations in tissue. Consequently, tumor areas without evident morphological changes can be detected, suggesting that a diagnosis could be obtained possibly at earlier stages.

6.2 Conclusions

6.2.1 Machine learning approaches in MALDI-MSI: clinical applications

This review is the result of a wide literature search in the field of MALDI-MSI, with particular emphasis on the application of machine learning approaches to MALDI-MSI data for clinical purposes.

The review describes the computational aspects of the statistical analysis of MSI data. In particular, the review starts with describing the data obtained from a MSI analysis and its multidimensionality. The review then details the preprocessing of the spectra in order to prepare the data before submitting it to the statistical analysis, by discarding the analytical variability and enhancing the biological information. Finally, the concepts of clustering, feature selection and classification are explained, and examples of applications of MALDI-MSI are reported.

Additionally, the review provides a tutorial to perform the aforementioned operations in freely available software (Orange Canvas and Weka).

The scope of the review was to evaluate the state-of-the-art algorithms that are used in the literature for MALDI-MSI applications in the clinical environment and, in particular, their employment for feature selection and classification. By addressing at the possible application of MALDI-MSI in the daily clinical routine, the results of the studies proposed in the review prove the capability of the MALDI-MSI technology in providing molecular information that can be exploited by machine learning algorithms to perform the aforementioned operations in order to provide assistance in the diagnostic process of particularly difficult cases.

The work performed for this review allowed the evaluation of the state of the art of the algorithms and their implementation, in order to provide the fundamentals for the development of a software that employs such algorithms for performing the clinical classification of specimens through MALDI-MSI data.

6.2.2 A Support Vector Machine Classification of Thyroid Bioptic Specimens Using MALDI-MSI Data

The results of the work proposed for this publication demonstrate the capability of MALDI-MSI data in classifying patients by employing machine learning algorithms. In particular, the obtained results paved the way for the following work, with the aim of exploiting the mass spectrometric data for supporting the pathologists in the daily diagnostic practice.

In this work, a first version of the software, written in the R environment, is proposed, in the form of R scripts without any graphical user interface.

MALDI-MSI data of thyroid cytological smears was imported and submitted to preprocessing, in order to yield a peaklist matrix for statistical analysis. The data underwent feature selection, by the employment of the *Recursive Feature Elimination* (RFE) algorithm which fit a *Partial Least Squares* (PLS) model to evaluate feature weights and discarded the least relevant features for classification. Out of 144 features, the 20 most informative ones were preserved and passed to a *Support Vector Machine* (SVM) (with *Radial Basis kernel function*) classification model for training. The model underwent tuning over a set of parameters in order to maximize its classification performances.

The results show that the feature selection step is mandatory in order to speed up the computational time and to increase the classification capability of the model. The reason behind this behavior is addressed as the “curse of dimensionality”: the presence of a high number of features, most of which are highly correlated and/or invariant, therefore redundant, makes algorithms less efficient and more prone to overfitting, due to their employment of many confounding features. Additionally, the feature selection procedure yields a restricted number of significant features, which can be further investigated under a biological point of view, to be employed as biomarkers for classification.

Finally, the application of the SVM model to clinical specimens of thyroid smears, also in the form of the proposed pixel-by-pixel classification, proves the capability of mass spectrometric data in possibly aiding the diagnostic procedure by providing a molecular insight of the specimen with a good discriminative power.

6.2.3 Combining multiple classifiers for the pixel-by-pixel classification of bioptic specimens by MALDI-MSI

The body of work present in this paper shows the evolution and improvement of the software that was initially proposed. The proposed software aims at providing the end user with a full data elaboration workflow, in order to exploit mass spectrometric data for the purpose of patient classification.

The software is split in three modules: the first allows to import and preprocess MALDI-MSI data according to different parameters, the second allows to perform feature selection and model training/tuning, while the third applies the trained model to the spectral classification of specimens. All the three software modules come with a simple and intuitive Graphical User Interface (GUI) to allow its use on a daily basis by end users.

The software makes use of a combination of several classifiers (*Support Vector Machines with Radial Basis kernel function*, *Support Vector Machines with Polynomial kernel function*, *Support Vector Machines with Linear kernel function*, *Partial Least Squares*, *Random Forest*, *Naive Bayes Classifier*, *k-Nearest Neighbor*) to perform classification. In particular, each classifier selects the best discriminatory features on the same training set, by employing the *Recursive Feature Elimination* (RFE) algorithm, and undergoes tuning over a set of parameter to maximize the performances. When applied to new mass spectrometric data, each classifier predicts the outcome of each spectrum of the MALDI-MSI dataset, performing a spectrum-by-spectrum (i.e. pixel-by-pixel) classification of the specimen. At the end, for each spectrum, the final outcome is determined as a weighed vote among classifiers, with the weight established as the reliability of the model in classifying new spectra (posterior probability according to the *Bayesian framework*).

The results obtained when applying the full software stack to data coming from the MALDI-MSI analysis of clinical specimens show that the software has the capability of highlighting the presence of sub-areas of tissue corresponding to tumor, therefore potentially providing aid in the diagnostic procedure. By exploiting the mass spectrometric data directly, the diagnosis is performed at the molecular level, possibly highlighting changes that are not necessarily correlated with morphological alterations.

6.2.4 Generating iMatrixSpray methods for MALDI-MSI analyses

The work proposed in this technical note will provide the scientific community with the possibility to easily generate G-code methods for the iMatrixSpray device, which, at the moment, does not come with a full-featured application that can allow the end user to tweak the spraying method parameters extensively. Additionally, the technical note will include method details for different applications: from spraying different matrices (e.g. sinapinic acid and CHCA) to spraying trypsin, for the analysis of different kinds of molecules, such as proteins, peptides, lipids and metabolites. All the methods will be generated through the software by the laboratory and provided to the scientific community, in order to have a reference starting point for the MALDI-MSI analysis of such molecules.

6.3 Future perspectives

The analysis of huge amount of data provided by new advanced mass spectrometric instrumentations (higher spatial and mass resolution) and the computational hardness of many data analysis and data mining procedures require more complex and efficient tools to exploit useful information and to provide adequate solutions.

6.3.1 Graph Theory

Graph theory allows to model several types of systems, both natural and human-made, ranging from biology to sociology science. In this context, a graph provides a system representation in terms of relationships among elements. More formally, such elements are represented by vertices and relations between vertices are represented by edges.

Graphs will be mainly used to provide alternative representations for the standard data cube and to model relationships between the spectral profiles. In particular, some sub-graph structures, typically applied to detect communities in social network analysis and patterns in biological networks, will be considered to indicate various types and configurations of “dense” groups (sub-graphs) such as cliques, which are sub-graphs where all the pairs of elements are connected by an edge. The MSI problem can be given as a graph, where vertices represent specific profiles at specific coordinates and edges represent correlations between profiles.

Since ideal tissue partitions should contain spectra with high degree of similarity, related to clusters of cells of the same molecular nature, potential segments of MS images can be provided by exploiting the relational information associated with the spectra from the corresponding sub-area. For example, by assuming that a correlation between spectral profiles of contiguous areas within a damaged tissue exists (again, for high degree of spectral similarity), correlation analysis should be exploited to provide graphs where nodes represent profiles at specific physical coordinates and edges identify significant higher correlations among the considered profiles.

This study will employ genetic algorithms applied to optimization problems. Genetic algorithms are inspired to the biological concept of natural selection, involving mutation, cross-over and selection events. They start by generating a population of chromosomes, encoded by 0 and 1 in most cases, representing a possible solution to the optimization problem. Each iteration, also called generation, returns a possible solution to the problem, evaluated by a function called fitness. At each generation, the chromosomes undergo mutation (change between 0 and 1) and cross-over (exchange of blocks of 0-1), occurring with a certain probability. The more fit individuals are selected and carried on towards the next generation. The algorithm terminates either when a maximum number of generations is obtained or when a certain threshold (e.g. minimum) of the fitness function has been reached. For example, genetic algorithms applied to feature selection encode features to retain with 1 and features to discard with 0; after mutation and cross-over events, the solution represented by the best result of the fitness function (e.g. discriminatory power of the features measured in terms of classification accuracy) is preserved at each iteration; at the end, the best solution, i.e. the best subset of features, is returned.

The process will start by applying a simple case, i.e. designing particular genetic operators for the considered problem. In this case, selection, mutation and crossover will be adapted and interpreted properly to provide consistent hypotheses for the “clique vs. independent set” partitioning. The resulting genetic algorithms will serve both to provide first tentative approximations for the MSI segmentation problem and, more in general, to design similar approaches for partition-based computational problems.

6.3.2 Distributed Machine Learning

The huge amount of data obtainable by advanced mass spectrometers and the intrinsic computational complexity of the respective algorithmic procedures make most of the standard approaches of predictive inference virtually impracticable. One task that should really be taken into account for implementation is an effective distribution of the computational costs over different virtual processors, e.g. using clouds or multi-core systems. In particular, the research, in this case, will be focused on how to obtain an efficient computational load distribution, by providing both parallel concept learning and a proper delivering of the training data on different nodes of the cluster, following the MapReduce paradigm.

The employment of one single classifier can be limiting for the reliable classification of specimens, even when algorithms particularly suitable for feature-rich data (such as Support Vector Machines) are chosen. Multiple classifiers can be applied simultaneously and combined (with a weight related to their reliability) for the prediction of the outcome of a patient, as described in Chapter 4). The computations can be distributed over a cluster, simply by mapping both the training and the test phases of the individual learning models into different nodes, and finally combining (i.e. reducing) the different responses (similarly to the standard “ensemble learning”) to get the definite target concept.

List of publications

Articles in Journals

- Smith A, Piga I, Galli M, Stella M, Denti V, Del Puppo M, Magni F. “**Matrix-Assisted Laser Desorption/Ionisation Mass Spectrometry Imaging in the Study of Gastric Cancer: A Mini Review**”. Int J Mol Sci. 2017 Dec 1;18(12). pii: E2588. doi: 10.3390/ijms18122588.

- Galli M , Pagni F, De Sio G, Smith A, Chinello C, Stella M, L'Imperio V, Manzoni M, Garancini M, Massimini D, Mosele N, Mauri G, Zoppis I, Magni F. “**Proteomic profiles of thyroid tumors by Mass Spectrometry-Imaging on Tissue Microarrays**”. Biochim Biophys Acta. 2016 Dec 8. pii: S1570-9639(16)30259-X. doi: 10.1016/j.bbapap.2016.11.020.

- Smith A, L'Imperio V, Ajello E, Ferrario F, Mosele N, Stella M, Galli M, Chinello C, Pieruzzi F, Spasovski G, Pagni F, Magni F. “**The putative role of MALDI-MSI in the study of Membranous Nephropathy**”. Biochim Biophys Acta. 2016 Nov 23. pii: S1570-9639(16)30250-3. doi:10.1016/j.bbapap.2016.11.013.

- Galli M , Zoppis I, Smith A, Magni F, Mauri G. “**Machine learning approaches in MALDI-MSI: clinical applications.**” Expert Rev Proteomics. 2016 Jul; 13(7):685-96. doi: 10.1080/14789450.2016.1200470.
- Galli M , Zoppis I, De Sio G, Chinello C, Pagni F, Magni F, and Mauri G. “**A Support Vector Machine Classification of Thyroid Bioptic Specimens Using MALDI-MSI Data**”. Adv Bioinformatics. 2016;2016:3791214. doi: 10.1155/2016/3791214.
- Pagni F, L'Imperio V, Bono F, Garancini M, Roversi G, De Sio G, Galli M, Smith AJ, Chinello C, Magni F. “**Proteome analysis in thyroid pathology**”. Expert Rev Proteomics. 2015 Aug; 12(4):375-90. doi: 10.1586/14789450.2015.1062369.
- De Sio G, Smith AJ, Galli M, Garancini M, Chinello C, Bono F, Pagni F, Magni F. “**A MALDI-Mass Spectrometry Imaging method applicable to different formalin-fixed paraffin-embedded human tissues**”. Mol Biosyst. 2015 Jun; 11(6): 1507-14. doi: 10.1039/c4mb00716f.
- Pagni F, Mainini V, Garancini M, Bono F, Vanzati A, Giardini V, Scardilli M, Goffredo P, Smith AJ, Galli M, De Sio G, Magni F. “**Proteomics for the diagnosis of thyroid**

lesions: preliminary report”. Cytopathology. 2015 Oct; 26(5):318-24. doi: 10.1111/cyt.12166.

Book chapter publications

- Mosele N, Smith A, Galli M, Pagni F, and Magni F. "**MALDI-MSI Analysis of Cytological Smears: The Study of Thyroid Cancer**". Imaging Mass Spectrometry Methods and Protocols. Methods in Molecular Biology, Springer. 2017. 1618;37-47.
- Smith A, Galli M, L'Imperio V, Pagni F, and Magni F. "**MALDI-MS Imaging in the study of Glomerulonephritis**". Imaging Mass Spectrometry Methods and Protocols. Methods in Molecular Biology, Springer. 2017. 1618;85-94.

Acknowledgements

This thesis was written using the font **IBM Plex Sans** (SIL Open Font License v1.1) and **LibreOffice** (v6.0.1.1) (**The Document Foundation**) (MPLv2.0 license), under the **Ubuntu GNU/LINUX** operating system, and saved as **Open Document Format** (.odt) file.

RESPONSES TO THE REVIEWERS

I would like to thank the reviewers for reading the thesis and providing criticism and suggestions, which have been positively taken and translated into changes to the thesis text accordingly.

REVIEWER 1

1) Figures often lack measurement units and it's difficult to grasp some details due to their small size.

I truly apologize for the quality of the pictures, which had to fit to the page format imposed by the thesis guidelines. I can provide full-resolution images upon request.

The apparent lack of measurement units on the spectra is due to the fact that mass-to-charge ratio (m/z) and arbitrary intensity (a.i.) are both axis labels and measurement units.

2) The description of the SVM techniques, occurring in several parts of the thesis, is described in terms of the maximization of distances between the closest observations. Now, the position of observations is fixed (at least in the original space), and what is maximized is the so-called margin of the separating surface.

The latter, in turn, is defined as the sum of the distances between the closest points for each class and the surface itself.
I agree with the reviewer that the definition of Support Vector Machines can be misleading and therefore sounds incorrect. The definition has been corrected in the introduction accordingly (p. 37).

3) There is some confusion in the definition of validation and test phases that need to be clarified: the first term (validation) commonly refers to the tuning of hyper-parameters of the inferred model (the so-called model selection involving for instance the trade-off parameter C in a SVM), while the second one (test) is used to denote the phase in which the generalization ability of an inferred model is assessed (e.g., using hold-out or cross-validation techniques).
The description of such phases has been extended and modified accordingly, and it should be more clear (p. 38).

4) There is also confusion w.r.t. the meaning of "cross-validation", which does not correspond to dividing the available data into two different sets devoted to training and testing, respectively. The latter procedure is commonly referred to as "hold-out". Cross-validation refers to a different validation *technique dividing data into k subsets having approximately the same size, and subsequently executing the learning process k times, each time using a different subset for testing and the remaining ones for training. The text should be modified accordingly.*

I agree with the reviewer on the fact that using the word “two” and not even mentioning *k* can be misleading: the definition has been corrected accordingly (p. 38).

5) The formula in the middle of p. 58 need a suitable introduction.

I strongly agree with the reviewer’s suggestion, however, according to the guidelines for writing the thesis, publications have to be reported as they are on the journal.

*6) The proposed normalization techniques (p. 58) do not make the data normally distributed; rather, they change a generic normal distribution into a standard normal one. Moreover, the 1-norm involves the sum of the *absolute values* of intensities.*

I strongly agree with the reviewer’s suggestion, however, according to the guidelines for writing the thesis, publications have to be reported as they are on the journal. Therefore, the suggested modifications have been taken into account in the modification of the text in the introduction section (p. 21).

7) A formal definition is needed for some of the introduced terms (e.g., tensor, overfitting, curse of dimensionality, ROC, PPV, homoscedastic, p-norm, node, linkage, ...).

I thank the reviewer for suggesting such modifications, which allow for a better and wider comprehension of the methods and the aims described in the thesis.

8) Clustering is introduced in p. 64 as a procedure which doesn't require any knowledge about data, although several of the used clustering algorithms do actually need some initial information (e.g., the number of clusters). The corresponding description should mention this fact.

I strongly agree with the reviewer's suggestion, however, according to the guidelines for writing the thesis, publications have to be reported as they are on the journal. Therefore, the suggested modifications have been taken into account in the modification of the text in the introduction section (p. 25).

9) The agglomerative nature of hierarchical clustering should be addressed, as well as the procedure translating a dendrogram

into a set of clusters. In this respect, some content of the caption of fig. 66 could be moved into main text. Moreover, there is some ambiguity in the joint use of "distance" and "similarity".
The text has been modified accordingly, to provide a more clear and detailed explanation of agglomerative hierarchical clustering (p. 25).

10) In some points it is stated that feature selection discards all noise, it is advisable to slightly modify these statements writing that this happens in principle.
I strongly agree with the reviewer on the fact that feature selection is supposed to discard the non-informative noisy variables from the data, but this happens only in principle, since, as stated throughout the thesis, it is a critical and delicate procedure. However, the occurrence mentioned by the reviewer happens to belong to a publication, therefore the original text is reported in the thesis (p. 77).

11) The criticism about the dependency of diagnoses on training and experience of pathologists could in principle move also the labeling process of the datasets used as input to ML procedures. Some comments about the robustness of the latter *(maybe in a footnote) are needed in order to give meaning to the criticism itself.*

I thank the reviewer for pointing out such a delicate aspect of machine learning in the training phase. The issue, being relevant to the work presented in the thesis, has been addressed in the Discussion, instead of a footnote (p. 167).

12) The possibility to build training and test sets in Orange Canvas according to the value of features (p. 87) should be motivated: usually examples are placed randomly in the two sets in order to ensure a fair subdivision.

I strongly agree with the reviewer's suggestion, however, according to the guidelines for writing the thesis, publications have to be reported as they are on the journal.

13) The supervised branch of ML is most of the time identified with the realm of classification problems, while regression is not generally mentioned.

I strongly agree with the reviewer on the fact that regression is not mentioned in the thesis as a supervised approach.

The text has been modified accordingly. However, being out of the scope of the thesis, the regression concept is addressed only briefly and partially (p. 37).

14) The proposed tutorial often lists parameters of learning algorithms without giving any explanation for them.

I strongly agree with the reviewer's suggestion, however, according to the guidelines for writing the thesis, publications have to be reported as they are on the journal.

15) In the five-year review of p. 91, the desiderata on hardware advancements should also mention GPUs.

I strongly agree with the reviewer's suggestion, however, according to the guidelines for writing the thesis, publications have to be reported as they are on the journal.

16) The choice of using only THY3 patients in the validation procedure (p. 128) should be motivated.

I strongly agree with the reviewer's suggestion, however, according to the guidelines for writing the thesis, publications have to be reported as they are on the journal.

The use of only THY3 patients for the validation phase is motivated by the fact that, in the clinical routine, diagnoses of the THY3 indeterminate reports is performed according to the cyto-morphological features learned from clear-cut diagnoses onto THY2-4-5: in the paper, we wanted to resemble the same situation by training classifiers onto clear-cut diagnoses and evaluating their classification capability onto indeterminate THY3 diagnoses.

17) It's difficult to give meaning to the sentence "the time taken by the process exponentially increases with the number of algorithms running" (p. 129): in the assumption that the times required by each of the considered algorithms is comparable, doubling the number of algorithms likely doubles the execution time, thus leading to a linear increase. Thus the sentence should either be modified or motivated in a convincing form.

I strongly agree with the reviewer’s suggestion, however, according to the guidelines for writing the thesis, publications have to be reported as they are on the journal.

18) Some comments of the results of experiments described in p. 130-131 are needed, highlighting that introducing feature selection turned a random classifier (ROC=0.5) into a meaningful one.

I strongly agree with the reviewer’s suggestion, however, according to the guidelines for writing the thesis, publications have to be reported as they are on the journal.

19) The use of "=" in the formula in p. 152 is wrong, because $P(d|h)P(h)$ equals the joint probability $P(h,d)$ and not the conditional one $P(h|d)$. To get to a correct formulation the "equal" symbol should be replaced by a "proportional to" one.

The suggested modification has been added to the text (p. 166).

20) The definition of clique (p. 192-193) is wrong: a clique is not a subgraph whose elements are connected by an edge. The text should be modified addressing to the key property that all pairs of elements in the subgraph be connected by an edge.

I thank the reviewer for the correction and modified the text accordingly.

21) Minor corrections:

- *remove comma after "Mass Spectrometry Imaging" (p. 29);*
- *in p. 31 the references [36-38] occur three times, consider a better organization of the text;*
- *the description of "Random Forests" in p. 37 refers to building a decision tree instead than a committee of decision trees;*

- _the implicit link in the caption of fig. 9 should become an explicit reference; moreover, the caption should be extended in order to describe the differences between the six graphs in the figure, as well as to introduce notation (e.g. used colors and shapes);_
- _the caption of fig. 14 does not describe the figure itself: nothing is said about the meaning of bullets, of their color, of the shown curve, of the role of the symbol γ; moreover "graphic" should be corrected into "graphical";_
- _remove comma after "can be obtained" (p. 39);_
- _reference [47] in page 47 should be reformatted;_
- _apparently there is some extra space between lines 14 and 15 in page 51;_
- _the citation of references is not consistent throughout the thesis;_
- _the sentence "the ability to achieve potentially every aim" should be reasonably softened;_
- _modify "statistic concepts" into "statistical concepts" (p. 55, but there are other occurrences in the text);_
- _the beginning of p. 63 should mention that preprocessing aims at discarding the analytical variability *due to noise*;_

- *change "the classification problem" into "a classification or a regression problem","For each," into "For each process,", "explained" into "provided", and "it exploits its intrinsic ... within the data" into "which exploits their intrinsic ... within them" (p. 64)*;
- *the SCiLS Lab 2014 software mentioned in Fig. 3 (p. 67) is not cited; the same applies to R (p. 127);*
- *"addressing at only" should be changed into "address only to", and "information preserved" into "preserved information" (p. 68);*
- *"K-Means algorithms" should be changed into "K-Means" (p. 74);*
- *"dependent from" should be changed into "dependent on" (p. 81);*
- *the sentence "and clear cell renal cell carcinoma cells" is unclear (p. 87); in the same page "classification problem solving" should be changed into "classification";*
- *change "any important features" either to "any important feature" or to "important features" (p. 89);*
- *Sect. 5 is missing in the section listing of p. 120;*
- *Table 1 in p. 122 is not referenced in the text;*
- *The symbol "//" in Table 2 is not explained (p. 130);*
- *Table 4 (p. 137) occurs after Table 5 (p. 131);*

- *There is an unmatched closed bracket at the beginning of p. 154;*
- *the symbol ">>>" in p. 155 has unknown meaning: probably it should be replaced by ">>";*
- *change "each one of them chosen" into "each chosen" (p. 158);*
- *change "relative short" into "relatively short" and "are enrolled" with "will be enrolled" (p. 166);*
- *change "initially proposed previously" into "initially proposed" (p. 188).*

All the minor revisions have been addressed and the modification were made in the text accordingly. Some of the suggested modifications, however, were proposed on already published articles: therefore, despite strongly agreeing with the reviewer's criticism, the original text was preserved in those instances.

REVIEWER 2

1) The machine learning section should be a bit extended, for example by giving some basic definitions: overfitting is not explained or discussed, and for example, genetic algorithms are not described.

In response also to the first reviewer, the machine learning section has been a little bit extended, with more extensive explanation of the concepts to aid the reader in a better comprehension of the thesis. Since genetic algorithms have not been implemented in the thesis and therefore can be a little out of the scope in the introduction, they have been described in the future prospectives section (p. 198).

2) *Consider also changing "Scope of the thesis" with "Thesis goals" or "Thesis contribution".*

The title "Scope of the thesis" was part of the guidelines for writing the DIMET PhD thesis, therefore it cannot be edited accordingly.

3) This chapter (Chapter 5) should be developed a bit (for example including an abstract and providing a more effective description of the contribution).

I thank the reviewer for suggesting such an improvement to the chapter. An abstract has been added, along with some more detailed motivation for the work and the contribution that has been derived from it. However, being a little bit far away from data analysis and closer to pure software development, but more related to the sample preparation, it can be misleading when compared with the rest of the thesis.