DOI:10.34918/85107

Wojciech SIKORA[1,*], Joanna POLAŃSKA[1]

Chapter 7. MASS CHANNEL SPATIAL DISTRIBUTION AS A TOOL FOR PEAK DETECTION AND ISOTOPE IDENTIFICATION IN MALDI TOF MSI DATA

7.1. Introduction

Mass spectrometry (MS) is a powerful tool widely used in medical research to inspect the molecular composition of samples such as biopsies from patients with suspicion of cancer. The identification of proteins, which are often drug targets and are present at low concentration levels in complicated mixtures, is significantly facilitated by the sensitivity of MS [1]. Information gathered by MS helps with diagnosis, early detection, and drug development not only for cancer but also for other diseases. This technique of molecular mass analysis exists for over a hundred years, and consists of three distinct steps: ionization, ion separation, and ion detection. In the first step, molecules of the sample are given an electric charge (ionization). This charge is used to differentiate between molecules of different masses in the next mass spectrometer module, the mass analyzer (separation). In the final step of MS analysis, the detector measures the quantity of each distinct ion (detection). This gives information about the relative abundance of each molecule in the sample. The result is a function of relative abundance and mass-to-charge (m/z) ratio of ions, called the mass spectrum.

Advances in soft ionization techniques like electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) and improved mass analyzers made it possible to detect higher-mass molecules such as proteins and peptides, which led to the wide use of MS in the field of Proteomics, [2] the study of proteins, their functions, and their compositions within biological structures.

[1] Department of Data Science and Engineering, Faculty of Automatic Control, Silesian University of Technology, Gliwice, Poland.
* Corresponding author: wojciech.sikora@polsl.pl.

Mass spectrometry imaging provides information about the spatial distribution of molecules at the surface of the sample. The general setup of the MSI experiment involves defining an (x, y) grid over the surface of the sample. The mass spectrometer then ionizes the molecules on the surface of the sample and collects the mass spectrum at each pixel on the sample, with the resulting spatial distribution defined by the pixel size [3]. Application of this technique to biological samples from a biopsy can provide information about the spatial distribution of molecules such as peptides and proteins, which in turn can be used to gain even more insight into the condition of the patient, causes, and course of the illness.

Matrix-assisted laser desorption/ionization (MALDI) is a key ionization technique in MS-based proteomics. The advantages of this ionization method are among others high-throughput, high-speed data acquisition, easy sample preparation, and imaging at a high spatial resolution [4]. All these properties make it possible to obtain a large amount of data in a short amount of time. Data that can provide useful insights if properly analyzed. The volume of the data generated by 2D imaging makes manual analysis infeasible therefore in recent years there has been a great increase in the usage of statistics and machine learning methods to analyze such data.

The paper aims to describe a full pipeline for analysis of MALDI time-of-flight (ToF) MSI data that uses the full potential of its data and produces a set of features that is nonredundant, small enough for use in various machine learning techniques, consists of features that can be connected to specific proteins and can be used to rapidly calculate values of the features given new records. The proposed method starts with the aggregation of all mass spectra into a single signal. Then the signal is transformed using a Gaussian filter and divided into small fragments, then considered separately. Data prepared in this way undergoes a feature extraction process where fragments of the aggregated signal are fitted with a gaussian mixture model (GMM) using the expectation-maximization (EM) algorithm, thus generating a representation of all data as a set of potential features defined by the position and the value of the area of its normal distribution. In the dimensionality reduction step, the set of normal distributions describing each fragment is compared using information about spatial distribution provided by MSI. Spatial distributions are compared using Peacock's test statistic [5] and merged if the value of this test statistic is below a data-driven threshold. Further reduction in dimensionality is done by noise filtering. The final step of this proposed pipeline is isoform search where isoforms of the same molecule are reduced to a single feature, once again by comparison of spatial distribution on the sample.

7.2. Analysis of MALDI MSI data

Analysis of MS data can be problematic, especially for MS methods that produce mass spectra for a wide range of m/z with lower resolution such as MALDI MS. With this resolution, ions of specific mass on the mass spectrum, give a gaussian-like shape called a peak. For samples with high complexity (many different molecules in the mixture) such as biopsy samples, those peaks can overlap making peak detection even more difficult. Moreover, molecules of the same mass can generate multiple peaks called isoforms as the result of receiving different charges in the ionization step of mass spectrometry. Finally, such signals are disturbed by many different sources of noise both low and high frequency.

Numerous techniques have been used for dimensionality reduction of MS data. In [6] genetic algorithm is used to iteratively select a small set of m/z values (channels) which using cluster analysis gives the best split of the data into two classes (cancer vs noncancer). A similar approach described in [7] used Artificial Neural Network with backpropagation to classify human tumors. Another heuristic method used for biological MS data feature selection is the Simulated Annealing Algorithm [8]. Exact approach commonly used for dimensionality reduction is Principal Component Analysis (PCA). In [9] PCA is used for dimensionality reduction of MS data, then clustered using Linear Discriminant Analysis. Yihui Liu in [10] uses discrete wavelet transform detail coefficients to extract features from the mass spectrum and then uses support vector machine (SVM) as a classifier. More recently [11, 12] used clustering with k-means based Divik algorithm. All of the above and other dimensionality reduction methods, including most filter, wrapper, and embedded methods of the classical knowledge discovery process, have been used to process and classify MS data. Nevertheless, none of these methods makes full use of the peak detection combined with spatial information that MSI provides.

7.3. Data

The data used for this paper comes from patients with head and neck cancer. A total of 4 samples have been used for acquiring the data. The samples were delineated by an expert pathologist with regions of epithelium and caner (Fig. 1). Each sample has a different size with a total number of over 150 thousand mass spectra. Each mass spectrum consists of over 100.000 m/z channels ranging from 800 DA to 4000 DA.

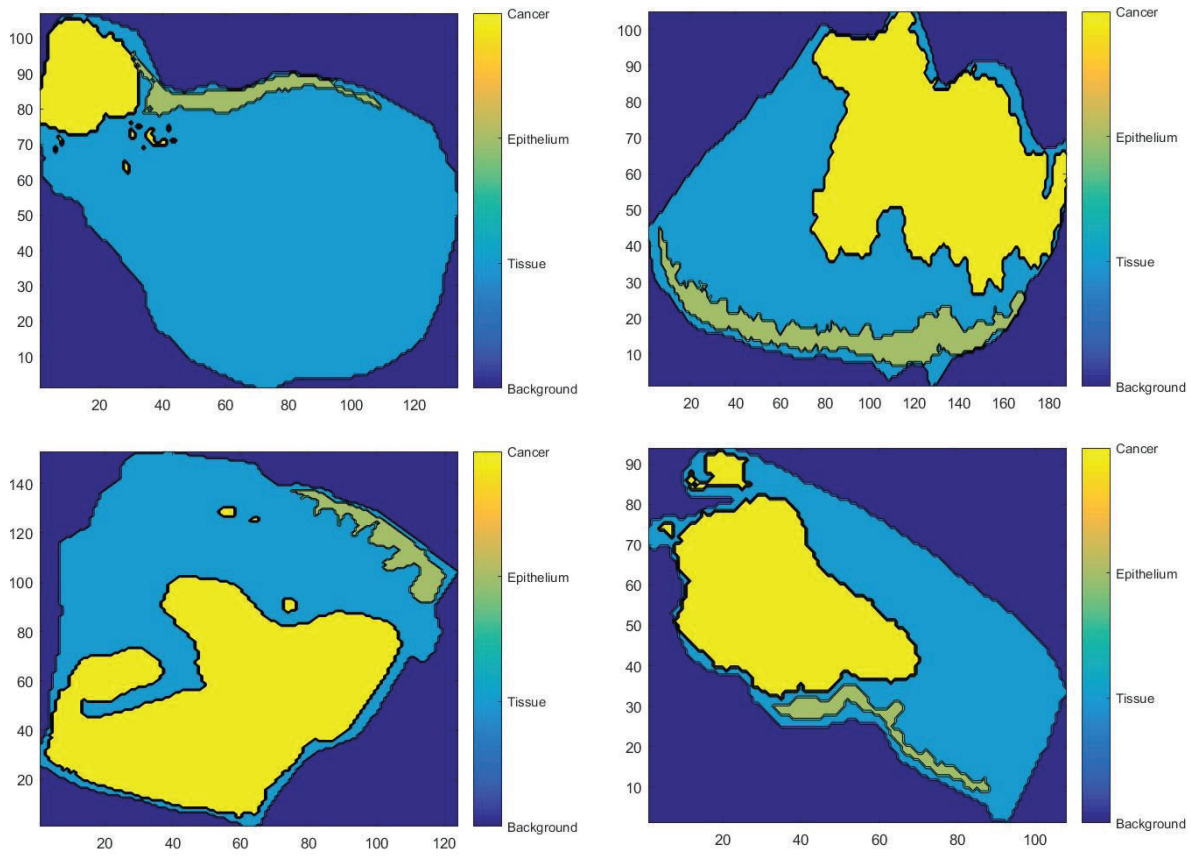

Fig. 1. Biopsy samples with marked epithelium and caner regions
Rys. 1. Próbki biopsji z zaznaczonymi obszarami nabłonka i nowotworu

7.4. Methods

The amount of computations necessary to compare the spatial distribution of two peaks on the sample and the need for a large number of such tests makes it impractical to apply the proposed dimensionality reduction process for each mass spectrum individually. Instead proposed solution is to combine all available data into a single signal and then try to identify peaks correlated with biomarkers on that representation of the data. To do that every value for each channel (m/z value) is aggregated.

7.4.1. Aggregation of mass spectra

The type of the aggregate has an impact on the remaining steps of the feature extraction and dimensionality reduction process. Using average, for example, makes the signal smooth with clear points of division between peaks, however, it can hide

meaningful information about potentially very important peaks that have high intensity only in a small number of mass spectra. (Fig. 2a). This problem is solved by choosing maximum as an aggregate, however, in this case, the signal is much rougher and further analysis is harder (Fig. 2b). Therefore intermediate solutions were tired, using the 95th percentile as an aggregate the trade-off between smoothness and information loss is more balanced. An interesting option is also combining the final results derived from each representation. The differences between the representation of all the data and an exemplary single mass spectrum can be clearly seen in the Fig. 2.

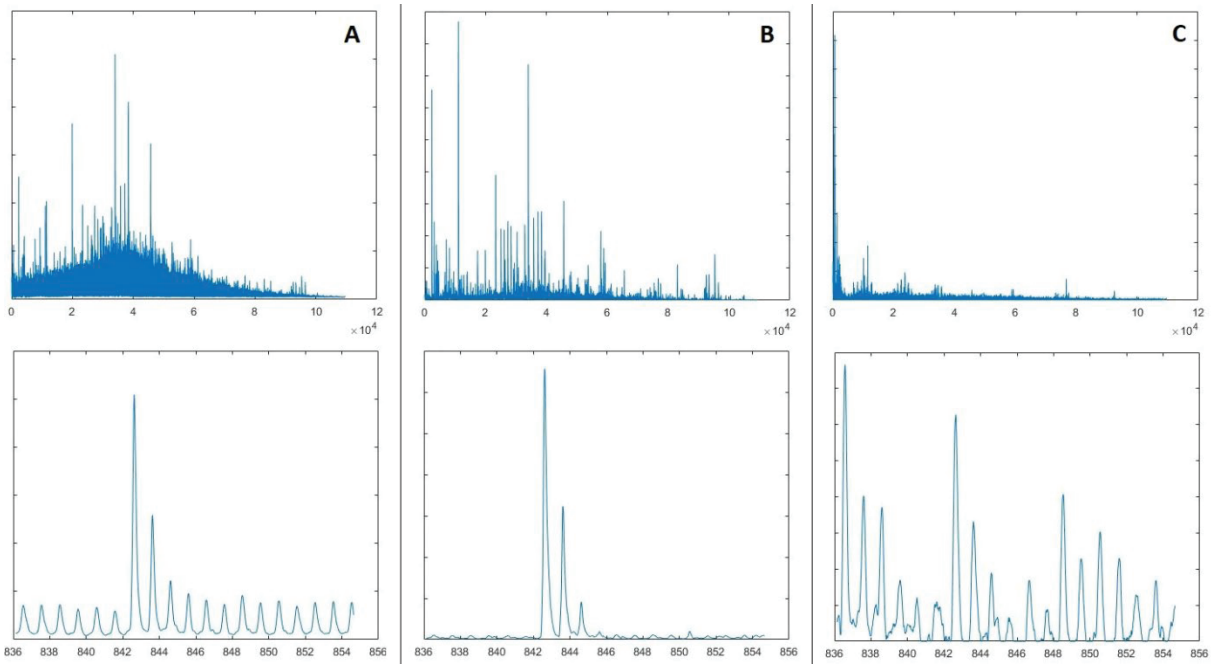

Fig. 2. Visualization of the entire signal (top) and small fragment (bottom) in average representation (A), maximum representation (B), and randomly selected single spectrum (C)

Rys. 2. Wizualizacja całego sygnału (góra) oraz jego niewielkiego fragmentu (dół) dla średniej (A), maksimum (B) oraz losowo wybranego pojedynczego spektrum (C)

Since molecules in MALDI MS are represented by gaussian-like shapes in the mass spectrum or a representation of all mass spectra, the next step of the proposed workflow is to represent the signal as a set of normal distributions around specific m/z values instead of as a list of intensities. This is done by dividing the signal into small parts and fitting them with a gaussian mixture model (GMM).

7.4.2. Signal division

Division of the signal is done by a simple local minima search. For the purpose of fitting GMM into parts, optimal points of division of the signal should be local minima that are relatively close to zero. Parts should also contain a small number of potential peaks ideally one peak or one set of overlapping peaks. Changing resolution of the mass spectra with the increase of m/z value can be problematic. In the lower m/z values, the difference between neighboring channel values is equal to 0.018 DA and it increases gradually to 0.041 DA at the end of the spectrum. In the higher values of m/z, the signal has also much more noise. The optimal process of the division was made by choosing the width of the moving window of the local minima search to be 120 channels. In order to find optimal points of division minima search was done on the signal first transformed with a gaussian filter of small width. This transform removed irregularities in the signal on the scale of a few neighboring channels leaving the shape of peaks unchanged. An additional restriction on the maximal value for the valley in minima search (restriction based on the local neighborhood) ensured that sets of overlapping peaks are not separated into additional fragments but considered together.

7.4.3. Modeling peaks

After splitting the signal into fragments with few potential peaks each, the parts are fitted with the Gaussian Mixture Model using the Expectation-Maximization algorithm in which a randomly selected set of starting conditions iteratively approaches to local optimum. In this case, the starting condition is a randomly selected GMM and the value being optimized is the log-likelihood of the model. Iterations of the EM algorithm are repeated until the shift between iterations, understood as the difference in log-likelihood of iterations is smaller than Ɛ represented as

$$\varepsilon = \left| \frac{L_{i-1} - L_i}{10^3} \right| \quad (1)$$

where:
$\mathcal{E}$ – threshold value for meaningful improvement in EM iteration,
L_i – log-likelihood of iteration i.

Because the EM algorithm is nondeterministic and highly dependent on starting conditions, the algorithm is run multiple times and the run with the best likelihood is chosen.

This process is then repeated with an increased number of elements of the initial GMM. The optimal number of normal distributions in the mixture is chosen by inspecting the gradient of BIC defined as

$$BIC = \log(n) * (3k - 1) - 2L \tag{2}$$

where:
BIC – Bayesian information criterion,
N – number of data points,
k – number of elements in GMM,
L – log-likelihood.

BIC is calculated each time and a decision based on the rate of the decrease of this value is made on whether to stop or add another element. Figure 3 shows some randomly selected parts with GMM fitted into them with this method.

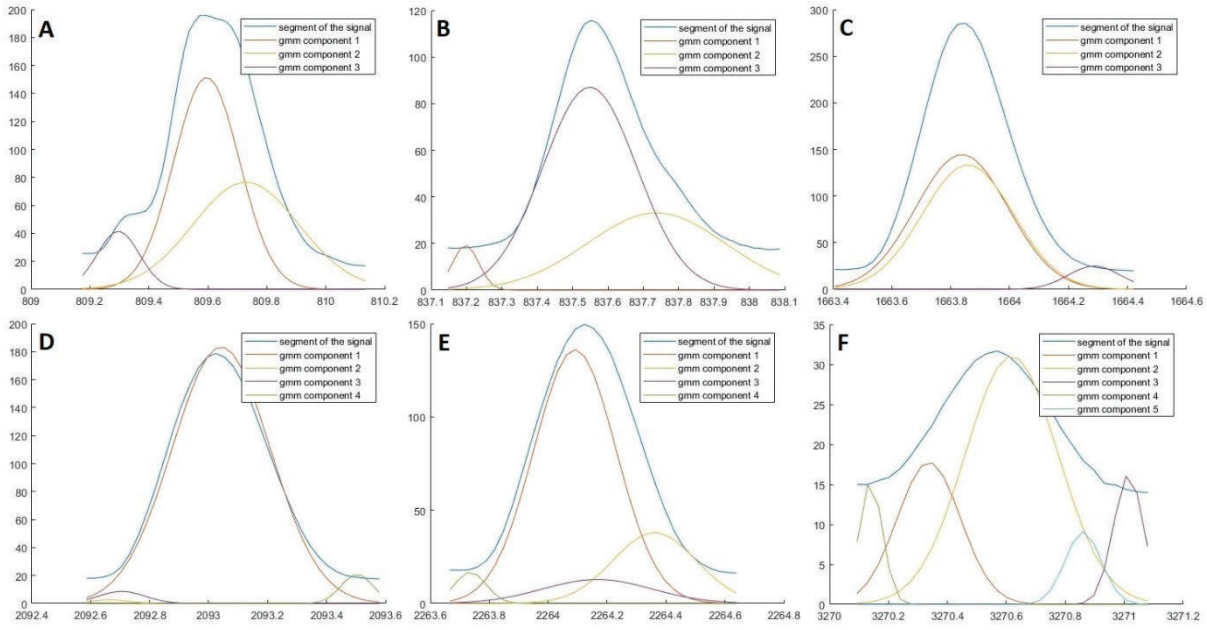

Fig. 3. Randomly chosen signal segments with fitted GMM
Rys. 3. Losowo wybrane segmenty sygnału z dopasowanym GMM

7.4.4. Spatial distribution based peak detection

At this point, all of the data is described by a set of normal distributions, identified by their position, sigma, and lambda values. Those distributions are potential features but many of them describe the same peak. Components 1 and 2 of GMM (Fig. 3c) are completely overlapping and therefore cannot describe different peaks. In Fig. 3b components 2 and 3 seem to be possibly part of a single peak. In Fig. 3a however, components 2 and 3, potentially describe two different, overlapping biomarkers with similar m/z values. In order to decide whether components are describing the same or different peaks, spatial distributions on samples are examined. In [5] Peacock describes a method for comparing spatial distribution by the extension of the Kolmogorov-Smirnov test into two dimensions. If two components within a segment are found to have the same spatial distribution on at least one of the samples, the component with a smaller area is removed from the set, and the area of the larger component is calculated as a sum of the two.

A statistical test used to decide whether two spatial distributions are the same is as mentioned the Peacock's test. However, for this test, no significance levels can be established [5], and therefore there is a need to define a threshold. In order to find the threshold value, more than 60.000 pairs of components were compared. The distribution of those values was fitted with GMM, and the threshold value for similarity of distributions was found (Fig. 4).

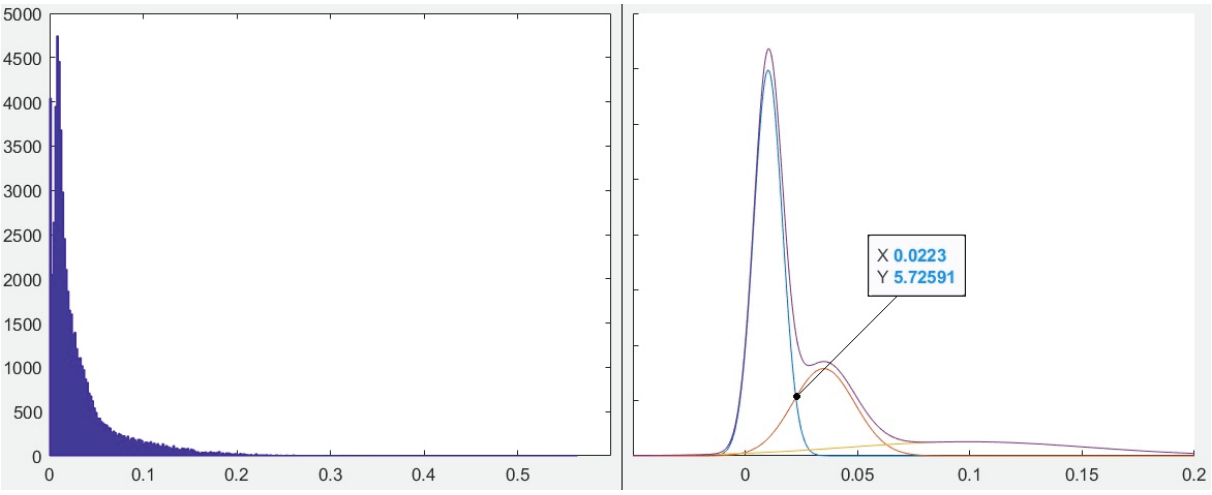

Fig. 4. Distribution of Peacock's test statistic and established threshold value for spatial distribution similarity

Rys. 4. Rozkład statystyki Peacock'a oraz wyznaczona wartość progowa dla podobieństwa rozkładu przestrzennego

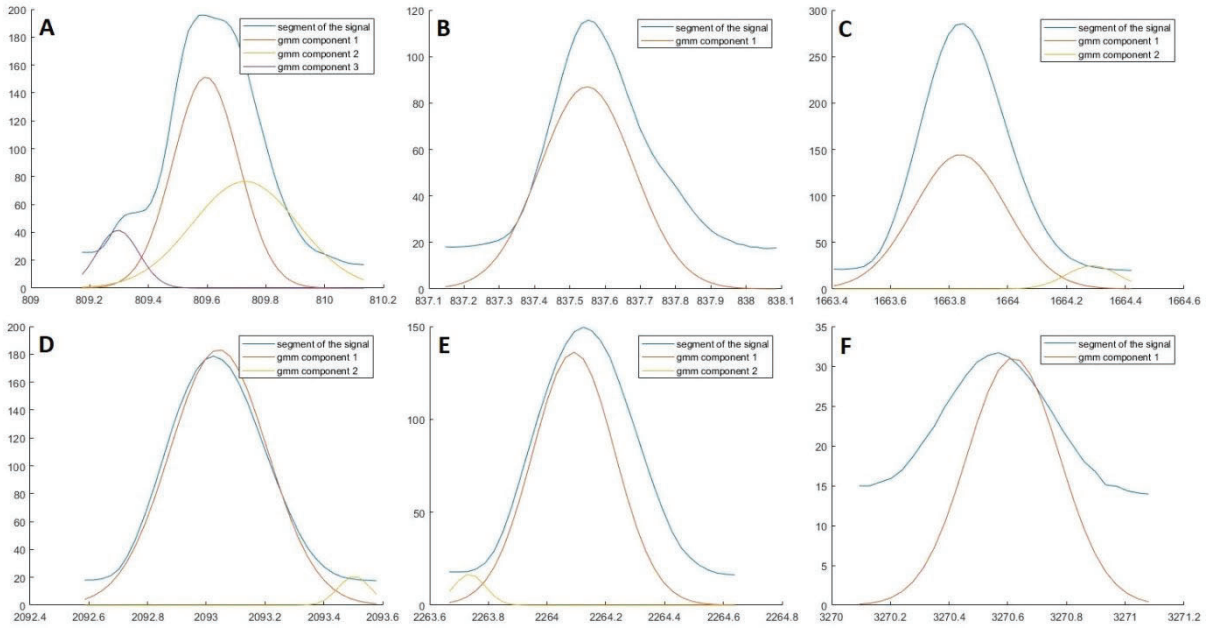

Fig. 5. Signal segments showed earlier with remaining GMM components after spatial distribution comparison

Rys. 5. Wcześniej pokazane segmenty z elementami GMM pozostałymi po porównaniu dystrybucji przestrzennej

Figure 5 shows the results of merging components with a similar spatial distribution. As predicted, components 1 and 2 from Fig. 3c were merged, as seen in Fig. 5c whereas GMM shown in Fig. 3a remains unchanged.

After comparing differences in the spatial distribution in segments shown earlier, we observe a clear reduction in their number (Fig. 5). Only Fig. 5a shows no difference in the number of GMM components. Components 1 and 2 from Fig. 3c were merged as predicted and Fig. 5f shows the resistance of this method to high-frequency noise characteristic of the high m/z part of the spectra. After merging, components associated with noise are also removed based on the signal-to-noise ratio. Finally, one more step is needed to complete the dimensionality reduction process. This step is the identification of isoforms.

7.4.5. Spatial distribution based isotope identification

Typically MALDI ionization results in singly ionized molecules, this, however, is not a rule, and therefore in the mass spectrum, molecules of the same mass can often be observed as a series of consecutive peaks. A series of peaks associated with isoforms of a single molecule is characterized by a 1DA distance between peaks and a similar shape (sigma in the context of normal distributions). The last step of the pipeline is

aimed at removing such isoforms and representing them in the final set of features as one feature, described by the localization of the main peak and the combined volume of all peaks.

In order to detect such isoforms, once again spatial distribution on the samples is considered. As it was mentioned, by the nature of isoforms, peaks of such series should be spaced evenly, with 1 DA of the distance between them on the mass spectrum, and should have similar sigma values. Therefore, when considering a peak as a potential start of isoform series, firstly the distance between peeks is considered, and then their shape. Once again thresholds for those values were calculated based on their distribution. A round-robin test with every peak within the 1 DA range through the entire signal would be very time-consuming so the narrowing of potential candidates for isoforms is key.

The search for isoforms starts from the first feature (with the smallest location value) and continues through the entire set of features. Firstly the condition of 1 DA distance is checked, once this condition is fulfilled the similarity of sigma values is checked. When the decision is made that currently considered features could be isoforms, based on those two conditions, the spatial distribution on the samples is evaluated. Figure 6 shows an example, where components 5 and 6 were preliminarily selected as possible isotopes by fulfilling distance and shape conditions but the spatial distribution shows statistically significant differences. Peacock's test statistic value is above the calculated threshold and therefore those features remain as separate features in the final set. Figure 7 shows an example of features for which Peacock's test statistic value was below the threshold, and were therefore considered as isoforms of the same molecule and merged.

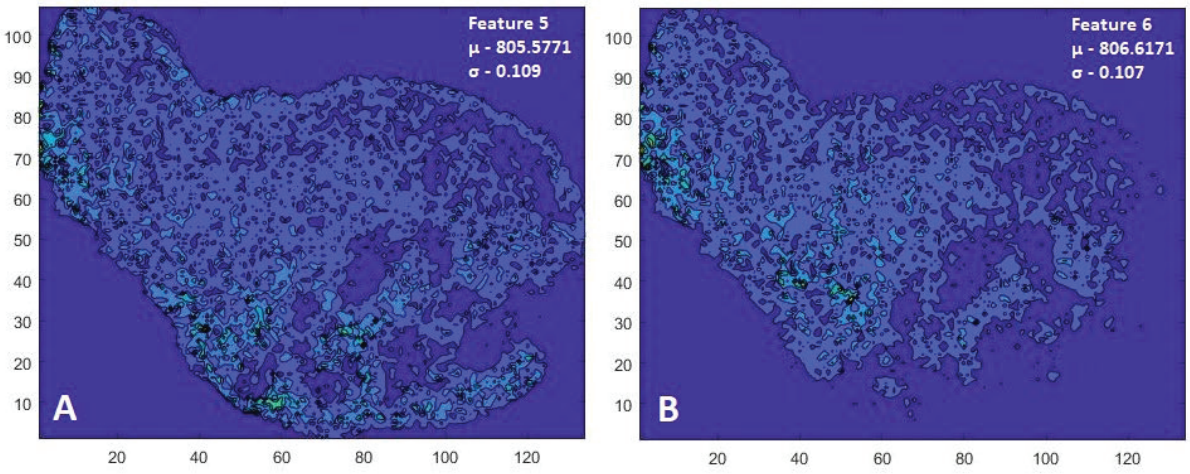

Fig. 6. Visualization of negative outcome of spatial distribution comparison between two features
Rys. 6. Wizualizacja negatywnego wyniku porównania dystrybucji przestrzennej pomiędzy dwoma cechami

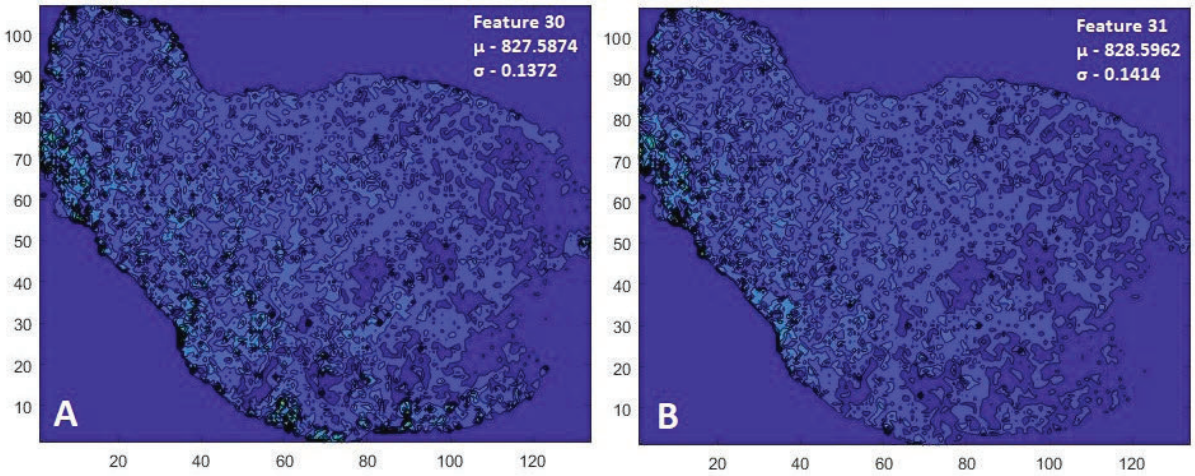

Fig. 7. Visualization of a positive outcome of spatial distribution comparison between two features
Rys. 7. Wizualizacja pozytywnego wyniku porównania dystrybucji przestrzennej pomiędzy dwoma cechami

One molecule can be represented by many isoforms, each separated by 1DA distance on the mass spectrum, so the search continues after the first match until no more matches are found.

7.5. Results and discussion

Computations of the workflo described in this paper,took around 48h to compute on, a computer with 32 GB RAM and i7-11700K 3,6G Hz processor. Considering the amount of the data and the large number of steps in the process, this time is reasonable for this proof of concept and once the features are defined, calculating values for the new data is a matter of seconds. The time needed for the completion of the full pipeline can, and will be optimized. Furthermore, with the usage of parallel computing, this value can decrease drastically, especially for the comparison of spatial distributions which is the most time-consuming step of the workflow.

Table 1

Number of elements in the dataset

Type of aggregation	Number of segments	Number of GMM components	Number of GMM components after merging	Number of GMM components after denoising	The final set of features
Average	3147	10811	4655	2684	1772
Maximum	2414	7939	3485	2060	1405
95th percentile	3131	10806	4662	2596	1629

Commenting on the results of feature extraction and dimensionality reduction, the 14 GB of data, describing tens of thousands of records, with tens of thousands of data points, were reduced to a manageable set of around two thousand features that take no more than 300 MB of memory. Table 1 describes how the number of dimensions of the data changed with each step. In this pipeline, the feature modeling is done using real-life properties of the investigated data, also feature selection and dimensionality reduction steps are based on the physical properties of the data. Thanks to this, we can expect that the final dataset consists of meaningful features, correlated with real properties of investigated data and that this set of features can be a good entry point for further analysis withfeature selection and machine learning.

Bibliography

1. G. Glish, R. Vachet: The basics of mass spectrometry in the twenty-first century, *Nature Reviews Drug Discovery* (2003) **2**:140–150.
2. K. Chughati, R.M. Heeren: Mass spectrometric imaging for biomedical tissue analysis (2010) **110**:3237–3277.
3. A.R. Buchberger, K. DeLaney, J. Johnson, L. Li: Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights, *Anal Chem.* (2018) **90(1)**:240–265.
4. R. Cramer, MALDI MS, *Methods in Molecular Biology* (2009) **564**:85–103.
5. J.A. Peacock: Two-dimensional goodness-of-fit testing in astronomy, *Monthly Notices of the Royal Astronomical Society* (1983) **202**:615–627.

6. E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta: Use of proteomic patterns in serum to identify ovarian cancer, *Lancet.* (2002) **359(9306)**:572–577.
7. G. Ball, S. Mian, F. Holding, R.O. Allibone, J. Lowe, S. Ali, G. Li, S. McCardle, I.O. Ellis, C. Creaser, R.C. Rees: An integrated approach utilizing artificial neural network and SELDI mass spectrometry for the classification of human tumors and rapid identification of potential biomarkers, *Bioinformatics* (2002) **18(3)**:395–404.
8. Y. Li, Y. Liu: A wrapper feature selection method based on simulated annealing algorithm for prostate protein mass spectrometry data, *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, (2008) 195–200.
9. R.H. Lilien, H. Farid, B.R. Donald: Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum, *Journal of Computational Biology* (2003) **10(6)**:925–946.
10. Y. Liu: Feature extraction and dimensionality reduction for mass spectrometry data. *Computers in biology and medicine* (2009) **39(9)**:818–823.
11. G. Mrukwa, J. Polańska, DiviK: Divisive intelligent K-means for hands-free unsupervised clustering in big biological data, *arXiv preprint* (2020).
12. K. Bednarzak, M. Gawin, M. Chekan, A. Kurczyk, G. Mrukwa, M. Pietrowska, J. Polańska, P. Widłak: Discrimination of normal oral mucosa from oral cancer by mass spectrometry imaging of proteins and lipids, *Journal of Molecular Histology* (2019) **50(1)**:1–10.

MASS CHANNEL SPATIAL DISTRIBUTION AS A TOOL FOR PEAK DETECTION AND ISOTOPE IDENTIFICATION IN MALDI TOF MSI DATA

Abstract

Mass spectrometry imaging provides information about biomarkers and their spatial distribution on the samples taken from patients. Analyzing this information can lead to advances in drug discovery and help with diagnosis. The rapid increase in the volume of the biological data gathered by mass spectrometry imaging made it necessary to develop new approaches for analysis. In this paper, we present a data-driven, hands-

-free approach for feature extraction and dimensionality reduction of MALDI ToF MSI data that can be used as the first step of the full machine learning analysis. In our approach, the feature modeling is done using the real-life properties of the investigated data. Also feature selection and dimensionality reduction are based on the physical properties of the data. Thanks to this we can expect that the final dataset consists of meaningful features correlated with real properties of investigated data. The result of the proposed pipeline is a nonredundant set of features small enough for further analysis for example with feature selection and machine learning.

Keywords: MALDI ToF MSI, feature extraction, peak modeling, dimensionality reduction, spatial distribution