Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie

Jugowice, 11th–15th September 2017

SPATIAL FEATURE SELECTION FOR FINDING BIOMARKERS USING IMAGING MASS SPECTROMETRY DATA

Michał Marczyk[1], Tomasz Smejkal[1]

[1]Silesian University of Technology
ul. Akademicka 2A, 44-100 Gliwice
[1]michal.marczyk@polsl.pl

ABSTRACT

Imaging mass spectrometry technique enables the combination of mass spectrometry measurements with spatial information on the tissue to be examined. This allows for the association of the molecular profile of the tissue with its morphological image. The technique is used to characterize the proteomic profiles of tissues, compare between different types of tissues (*e.g.*, tumor tissue and normal tissue), and to study tissue structure. Due to the high complexity of the measured signals it is necessary to find only the most representative spectral features (proteins or peptides), that could be good candidates for biomarkers, by examination of the spatial structure of the individual features. In this work an algorithm for efficient selection of only the most important spatially structured features is proposed. The algorithm is based on mixture modeling of mass spectrometry signal to define spectral features, application of two methods for ranking spectral features by their level of spatial structure and integration of the results obtained with these two methods. By analysis of biological tissue sample from patient with oral squamous cell carcinoma, it was proved that reduction of mass spectrometry signal to only important spectral features and integrating information from two spatial structure measures can enhance the potential for finding protein biomarkers by imaging mass spectrometry.

INTRODUCTION

Nowadays, we are looking for better solutions for the rapid diagnosis of diseases, especially cancer, where the time of detection and treatment is the most important issue [1]. Imaging Mass Spectrometry (IMS) based on matrix assisted desorption ionization (MALDI) technique allows the analysis of proteins from intact tissues, through visualization of their significant amounts in some regions. The tissue under investigation is divided into equally sized sections by the chessboard pattern and for each section the protein composition is obtained and stored as a single mass spectrum (Figure 1)[2]. Two-dimensional ion intensity maps can be created by plotting the intensities of a protein with given mass obtained as a function of its coordinates. The resulting images allow comparison of molecular distributions between different regions of the sample as well as between the samples [3]. Most importantly, resulting molecular images could be overlapped with morphological structures, allowing correlation between tissue structures and molecular features. The availability of tumor tissue samples that could be analyzed by IMS

(biopsies, resected tumors and xenograft models) increases the practicability of this approach in cancer medicine. It has potential to identify novel proteins, that were not obvious disease-specific molecules, but are involved in development of disease and interactions between surrounding tissue. The protein domain is directly affected in disease state, so proteomics holds special promise for biomarker discovery and MALDI-IMS isone of most promising approaches [4]. Protein biomarkers can be used clinically for screening, diagnosis or monitoring the activity of disease.

In a single spectrum composed of even hundreds of thousands of measurement points the most important elements are spectral signal peaks. It is assumed that each peak corresponds to a certain protein/peptide present in the analyzed samples. In high-resolution real data the shape of a single peak is right-skewed, so standard procedures for peak detection and quantification may not be accurate. In this work an efficient approach to computational processing of proteomic mass spectra is used [5]. The algorithm is based on modeling spectra by a mixture of Gaussian distribution functions and quantifying each peak to define spectral features (peaks) by a sum of the area under model components. Next, using the different measures for spatial structure on ion intensity maps, spectral features are graded. At last features are categorized by their importance and by integrating the results of clustering the most important sub-group of spectral features is selected.

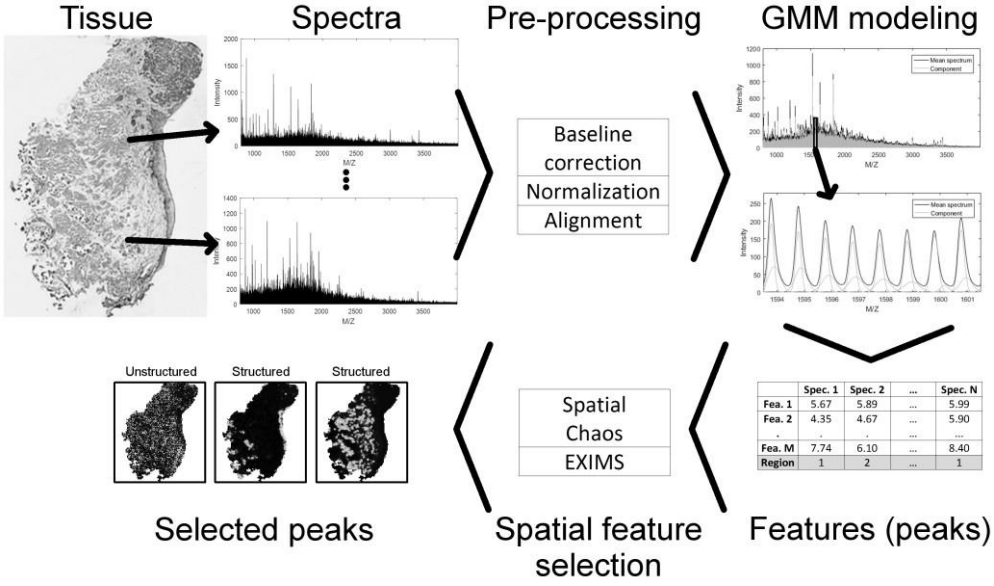

Figure 1. Diagram of collection and analysis of imaging mass spectrometry data for selection of spatially structured peaks.

MATERIAL AND METHODS

The tissue sample from a patient who underwent surgery because of oral squamous cell carcinoma was analyzed (Preparation_1 from [6]). The sample taken from fresh postoperative material was annotated by an experienced pathologist. Five tissue regions were distinguished: tumor area, normal epithelium, muscle, salivary gland and other structures. Mass spectra were recorded with the use of MALDI-TOF ultrafleXtreme mass spectrometer within M/Z range of 800-4000 Da.

The diagram of imaging mass spectrometry data analysis is presented on Figure 1. The raw spectra were resampled to a common M/Z space of 109 568 points, preserving their original distribution. For the set of all mass spectra the following pre-processing steps were performed:

baseline correction by estimating baseline within multiple shifted windows and spline approximation, spectral alignment with use of modified PAFFT algorithm and Total Ion Current normalization [7]. For all M/Z points intensity values across all spectra were averaged, creating mean spectrum, which was partitioned using the idea of "splitters" [5]. Particularly, each spectrum fragment was decomposed by using Gaussian mixture modeling and individual decompositions were aggregated to form the final mixture model. Too wide or too low components were filtered out [7]. In the last step, close components were merged to define the spectral features. The abundance of the spectral features was calculated as a sum of the area under the model components that belong to the particular spectral feature.

Currently, in literature there are two methods to measure spatial structure of ion intensity maps defined for spectral features, that was introduced for IMS data analysis. The first method is called spatial chaos (SC). Spatial chaos is defined as a lack of spatial pattern in the pixels intensities [8]. For each spectral feature the ion intensity map is created and a two-step edge detection filter for noisy images is applied to detect signal intensity edges. Next, a one-nearest neighbor graph on edge pixels is build. The measure of chaos is calculated on mean length of the graph edges. The value of the spatial chaos is low for an image exhibiting spatially structured intensity pattern and it is high for an image with spatially chaotic pixels intensities. Second method for identifying spectral features with structured spatial distributions was introduced as a part of data analysis pipeline for MALDIIMS data [9] called EXIMS. For each spectral feature the ion intensity map is created and preprocessed by median filtering and histogram equalization. The spatial structure is captured by calculating improved version of Gray level Co-Occurrence (GCO) matrix. The final measure constructed on selected values from GCO matrix should be high for structured images of spectral features and low for unstructured images. To provide only the most important spectral features kmeans algorithm was performed on values given by two spatial structure measures separately. The Box-Cox transformation to normality was made before clustering. The optimal number of clusters k was found using Dunn index.

With Anderson-Darling test non-normal data distributions were found for most spectral features. Due to non-normal data distributions the abundance of spectral features between regions defined by an experienced pathologist was compared using Kruskal-Wallis (KW) test with Nemenyi post-hoc procedure. Higher value of KW statistic brings higher statistical significance of given feature. The strength of dependence between variables was measured using Spearman rank correlation coefficient. In all tests the significance level was set to 0.05.

RESULTS AND DISCUSSION

MALDI-IMS analysis of one tissue sample gave measurements for 9 495 tissue sections, where each of them is represented by a single mass spectrum. First, the image sections located outside the tissue sample were neglected. The remaining 7676 spectra were assigned to 5 regions defined by pathologist, namely: tumor area (846 spectra), normal epithelium (359 spectra), muscle (1910 spectra), salivary gland (1145 spectra) and other structures (3416 spectra). By averaging the signal intensities over all spectra, mean spectrum was created. Constructed mixture model of mean spectrum consists of 6854 components with defined location, spread and weight parameters. 304 high variance and 73 low intensity Gaussian components were removed. Due to right-skewed shape of peaks most of them are modeled by two or more Gaussian components. By merging components close to each other (by investigating their location parameter), 3624 spectral features were defined. Each spectral feature correspond to protein or peptide present in the analyzed tissue. The abundance of each spectral feature was calculated by summing the area under the components associated with given spectral feature in each spectrum.

To grade the spatial structure of each spectral feature two measures were calculated: EXIMS and SC. EXIMS is based on non-parametric statistical measures, thus it is robust to diverse

distribution of intensities in analyzed ion intensity maps. SC algorithm gave different values for the spectral features abundance and the logarithmic transform of abundance, that was performed to reduce the skewness. In further text SC represents Spatial Chaos measure on spectral abundance in original scale and SC log represents Spatial Chaos measure on spectral abundance after logarithmic transformation. Furthermore, for 123 spectral components the value of Spatial Chaos could not be calculated by the software given from the authors, and they were neglected in all comparisons. The basic idea for introducing SC and EXIMS methods was the same: grade the spatial structure of the ion intensity map given by some spectral feature. However, the values obtained by using both methods are not highly correlated (correlation between SC and EXIMS equals to -0.21 and between SC log and EXIMS equals to -0.34). So, these two methods grade the spatial structure of given spectral feature in a different way. For each method the correlation between calculated measure of spatial structure and KW test statistic (that measures differences in abundance between tissue regions defined by pathologist) was calculated (Table 1). The highest correlation is observed for EXIMS measure, which suggest that it is the best method for finding spectral features with different abundance between tissue regions. The value of correlation coefficient for SC and SC log was about 3 times smaller than for EXIMS.

Table 1. Spearman rank correlation between spatial structure measures and Kruskal-Wallis (KW) test statistic.

Measure	**SC**	**SC log**	**EXIMS**
KW statistic	-0.22	-0.18	0.66

The goal of protein biomarker discovery is to find only the most important spectral features from the group of all features defined. Thus, k-means algorithm was made on values given from spatial structure measures separately to find the optimal cut-off value . In each case, by using Dunn index, three clusters were identified (Fig. 2). For EXIMS method the spectral features with the spatial structure measure higher than 1.903 (the cluster with the highest values of the EXIMS measure) were selected (863 spectral features). For SC methods the spectral features corresponding to the cluster with the lowest values of the SC measure were selected. The obtained thresholds were equal to 0.00158 for SC (341 spectral features) and 0.00071 for SC log (825 spectral features).

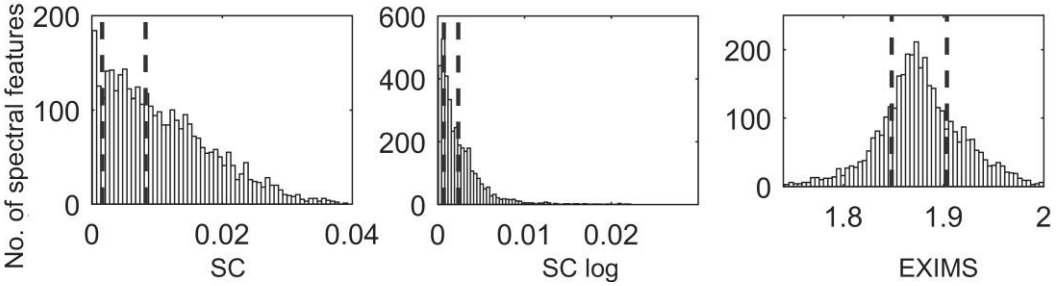

Figure 2. Distribution of different spatial structure measures. Dotted vertical lines represent thresholds found by k-means clustering.

Two measures of spatial structure give diverse classification of spectral features. It can be seen by visual inspection of scatter plots presented in Fig. 3. In this work it is proposed to integrate the results of EXIMS and SC by selecting the spectral features that are classified as the most important features for both methods. Particularly, SC+EXIMS corresponds to spectral features selected by EXIMS and SC methods (131 spectral features) and SC log+EXIMS corresponds to spectral

features selected by EXIMS and SC log methods (358 spectral features). Visually, the spectral features presented in the upper-left corner of images presented in Fig. 3 are selected.

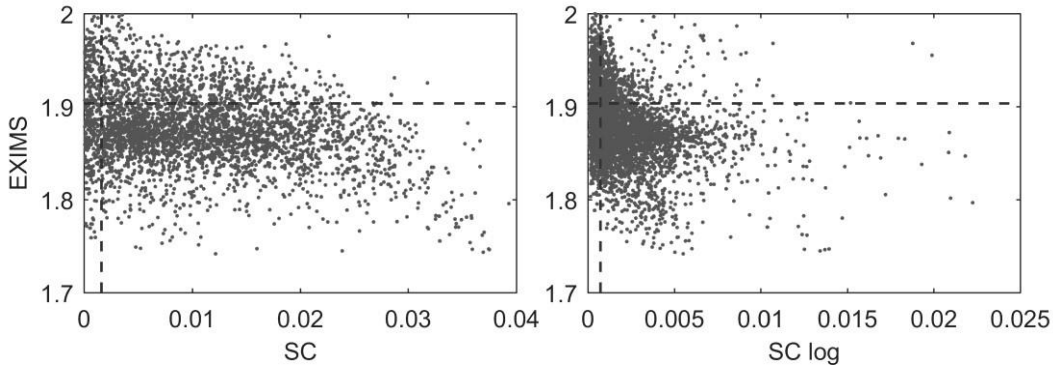

Figure 3. Association between Spatial Chaos (SC and SC log) and EXIMS measures. Dotted vertical and horizontal lines represent thresholds found by k-means clustering to distinguish only the most important features.

The statistical significance of spectral features selected by any method may be compared by presenting the distribution of Kruskal-Wallis test statistic among methods (Fig. 4). Also, the statistical difference between methods was obtained using again Kruskal-Walis test with Nemenyi post-hoc analysis. When all 3624 spectral features are taken into account, the median value of KW statistic is the lowest. Selecting only important spectral features gives increase in the median value of KW statistic and decrease in its range. The differences between the cases with and without feature selection are statistically significant. In average, EXIMS method gives more important features than SC and SC log. When we combine the results of EXIMS and SC, the increase in median KW statistic is even higher in comparison to other methods. No statistical difference was observed between SC+EXIMS and SC log+EXIMS methods.

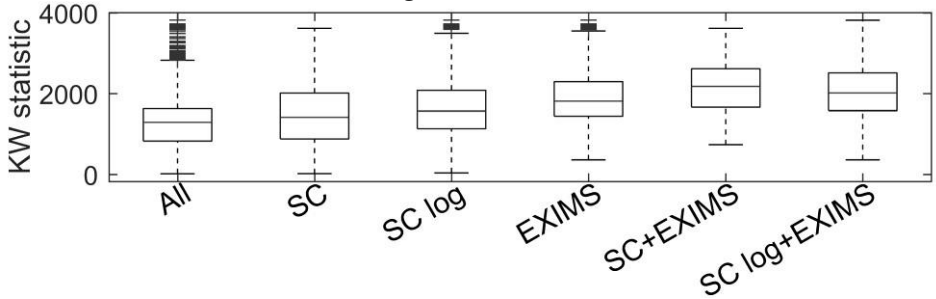

Figure 4. Comparison of Kruskal-Wallis test statistic between methods. All – all spectra features, SC,SC log and EXIMS – spectra features defined by given measure and k-means based thresholding, SC+EXIMS, SC log+EXIMS – spectral features defined by results integration.

In Fig. 5 some examples of ion intensity maps for spectral features selected by different methods are presented. The annotation for tissue regions made by pathologist is showed on the right side of Fig. 5. The selected spectral features corresponding to SC+EXIMS method, EXIMS only, SC only and other group was presented in a, b, c and d sub-images, respectively. These 4 spectral features were selected to visualize the general conclusions made on the analysis performed on all features associated with given methods. Spectral feature selected by integration of SC and EXIMS results shows high difference in abundance between epithelium plus cancer region and

other regions. In this group there are features that are visually the most similar to annotation made by pathologist, however there are only few features with abundance specific to given region. These results are similar to the one obtained in [6]. Spectral feature selected only by EXIMS method shows higher abundance in cancer region, with possible recurrence sites to other sections of analyzed tissue. In general, EXIMS gives wider regions with comparable spatial structure. In our analysis, SC method tends to select spectral features with smaller structured regions than EXIMS method. When no selection of most important features was made (Fig. 5 d), no spatial structure can be observed.

CONCLUSIONS

Imaging mass spectrometry is a powerful measurement technique that can be used to discover novel potential biomarkers. However, due to the high number of obtained spectra with noisy signal two important steps are required: (i) reduction of spectral signal to peaks, which was done using Gaussian mixture model based method, (ii) selecting group of only the most important spectral features, which was done by integrating the results from SC and EXIMS method. By analysis of real, biological sample tissue it was proven that proposed algorithm gives much better results than the existing methods performed separately and when no selection is performed. It is planned to improve the results of SC method by its modification and analyze the results of applying different clustering techniques. Also, the conclusions made based on the analysis of tissue from one patient must be verified using other tissue samples (only preliminary analysis was made).

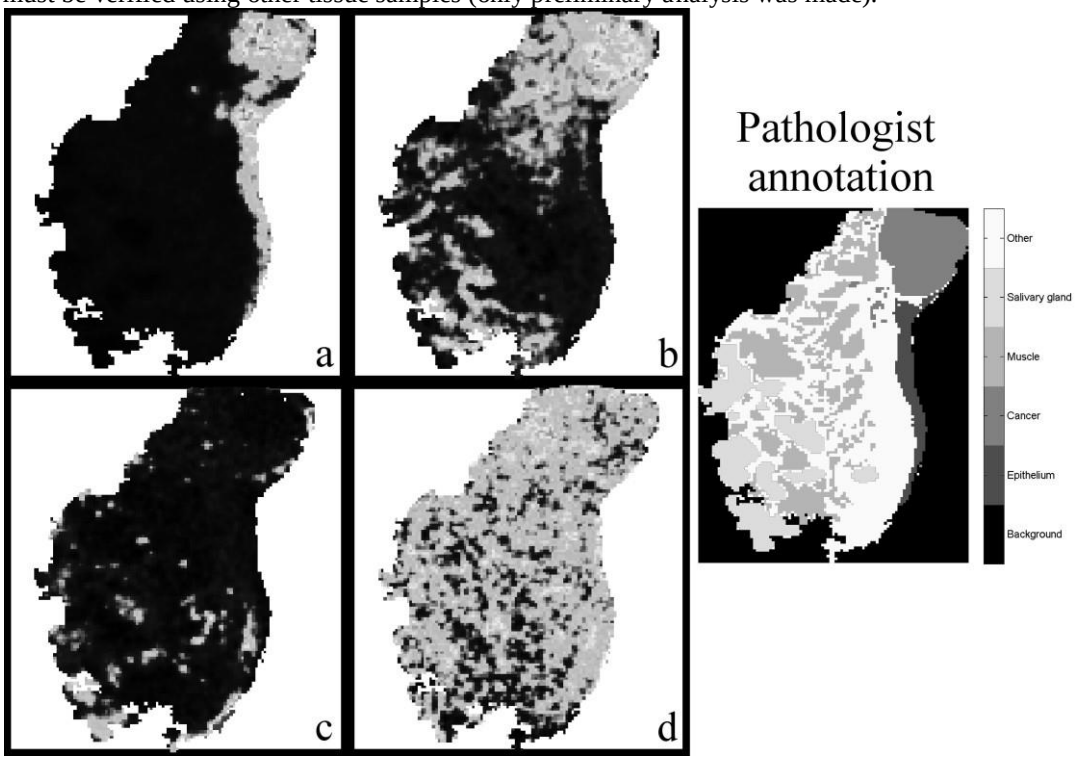

Figure 5. Examples of ion intensity maps for spectral features selected by the following methods: a – SC + EXIMS, b – EXIMS only, c – SC only, d – other features. On the right side there is a pseudo colored image of tissue annotation by an experienced pathologist.

ACKNOWLEDGMENTS

We are very thankful to prof. Piotr Widlak and dr Monika Pietrowska for access to the data and helpful discussions. This work was supported by NCN project number 2015/19/B/ST6/01736. All calculations were carried out using computer infrastructure funded by GeCONiI project number POIG.02.03.01-24-099/13.

REFERENCES

[1] M. Pernemalm and J. Lehtio: *Mass spectrometry-based plasma proteomics: state of the art and future outlook*, Expert Rev Proteomics **11** (2014), 431-48.

[2] R.L. Caldwell and R.M. Caprioli: *Tissue profiling by mass spectrometry: a review of methodology and applications*, Mol Cell Proteomics **4** (2005), 394-401.

[3] M.L. Reyzer and R.M. Caprioli: *MALDI-MS-based imaging of small molecules and proteins in tissues*, Current Opinion in Chemical Biology **11** (2007), 29-35.

[4] A.J. Scott, J.W. Jones, C.M. Orschell, T.J. MacVittie, M.A. Kane, and R.K. Ernst: *Mass Spectrometry Imaging Enriches Biomarker Discovery Approaches with Candidate Mapping*, Health physics **106** (2014), 120-128.

[5] A. Polanski, M. Marczyk, M. Pietrowska, P. Widlak, and J. Polanska: *Signal Partitioning Algorithm for Highly Efficient Gaussian Mixture Modeling in Mass Spectrometry*, PLoS One **10** (2015), e0134256.

[6] P. Widlak, G. Mrukwa, M. Kalinowska, M. Pietrowska, M. Chekan, J. Wierzgon, M. Gawin, G. Drazek, and J. Polanska: *Detection of molecular signatures of oral squamous cell carcinoma and normal epithelium – application of a novel methodology for unsupervised segmentation of imaging mass spectrometry data*, Proteomics **16** (2016), 1613-1621.

[7] M. Marczyk, G. Drazek, M. Pietrowska, P. Widlak, J. Polanska, and A. Polanski: *Modeling of Imaging Mass Spectrometry Data and Testing by Permutation for Biomarkers Discovery in Tissues*, Procedia Computer Science **51** (2015), 693-702.

[8] T. Alexandrov and A. Bartels: *Testing for presence of known and unknown molecules in imaging mass spectrometry*, Bioinformatics **29** (2013), 2335-2342.

[9] C.D. Wijetunge, I. Saeed, B.A. Boughton, J.M. Spraggins, R.M. Caprioli, A. Bacic, U. Roessner, and S.K. Halgamuge: *EXIMS: an improved data analysis pipeline based on a new peak picking method for EXploring Imaging Mass Spectrometry data*, Bioinformatics **31** (2015), 3198-3206.