1P

UDC 65.011

Małgorzata Plechawska-Wójcik
Lublin University of Technology
gosiap@cs.pollub.pl

Comprehensive Methods of MALDI-TOF Mass Spectrometry Data Analysis

The article presents a review of methods of MALDI-TOF data. There are many steps of mass spectrometry data analysis. It is complex task and it should cover several platforms. It is important to do comprehensive analysis to obtain useful results. The analysis should cover preprocessing and signals analysis, databases searching and peaks classification.

Introduction

Development of computational methods of mass spectral data processing is very expansive. Extensive literature in this field was created in the last few years. Research was conducted on such issues as: baseline removal, normalization, denoising, peaks detection and alignment.

Mass spectrometry (MS) is an analytical technique which performs chemical fragmentation of a sample into ions. Its basic is a measurement of mass to charge (m/z). This technique is used in such fields as: determining the structure of compounds, identification of compounds and mixtures of compounds, determining isotopic composition of elements in a compound. In practice, proteomic patterns may be used for early diagnosis, monitoring disease progression or effects of treatments [1].

One of the most popular MS technique used in proteomics is MALDI-TOF. MALDI-TOF [4] instruments employ a matrix-assisted laser desorption and ionization (MALDI) ion source as well as a time-of-flight (TOF) detection system. Samples are inserted in the vacuum chamber on a metal plate [1]. Afterwards they are ionized by a laser. Ions are accelerated by electric field until they reach a detector. Intensity is actual number of ions. Velocity achieved by ions in a drift tube is proportional to mass-to-charge ratio.

Obtained with this technique sample contains tens (or even more) of spectra. Each spectrum represents thousands of intensity measures with unknown number of protein peaks [2]. The challenge is to determine those peaks and align them between different spectra. Determined peaks serve as a futures in further processing tasks [3]. Interpretation of mass profiles is that each peak of the spectrum has its reflection in the composition of analyzed sample. It contains information which may be important for analyzed process and may help to find a meaningful biological conclusion.

Analysis of mass spectrometry data is a complex task. A process of gaining biological information and knowledge from row data is composed of several steps. All those steps need to be performed to get the information which may occur helpful in diagnosis or medical treatment tasks.

Using proteomic techniques as a way to support early diagnosing of diseases is an opportunity for developing of new way of treatment. There is a group of diseases which needs for new treatment and diagnosis approaches. For them typical ambulatory methods are not always useful. Particularly, the group contains a whole subgroup of cancer diseases.

Classification is essential part of mass spectrometry data. The most common classification task is based on supervised learning and it consists in categorizing data into two or more groups. It is possible to distinguish between ill patients and healthy donors or to check reactions (positive or negative) on the medical treatment. There is also possible to look for a stage of diseases progression.

Mass spectrometry data are characterized with high dimensionality. The number of observations is significantly lower than the number of features. Each patient has several thousand of data points or even more. Those data must be processed and dimension reduction techniques should be applied. This task determines success of the classification because of specificity of mass spectra data.

Classified objects are usually represented by vectors of observed, measured or calculated features. Supervised learning classification assumes, that there unknown function Φ is to be assigned to each object of population O as a label of one class. Classification process is based on the learning set U which is a subset of the whole data set O. Each element o_i of the learning set is composed of the object representation and a class label. This object representation is observation vector of features. The whole set is divided into c separated subsets and one subset observations are numbered among one of c classes. Such supervised learning is widely used in biomedical applications.

Mass spectra processing

Process of proteomic spectra analysis is composed of few steps. The most important are: calibration, binning, interpolation, trimming, denoising, baseline correction, normalization, peak detection and quantification as well as peak alignment.

Process of calibration consists in mapping gained time of flight values into mass-to-charge ratio. It is done by quadratic transformation [1]. Operation of binning with defined mask allows to reduce number of data points. A mask of 2 for example decreases size of sample by a half. Additionally, this process gives a noise reduction.

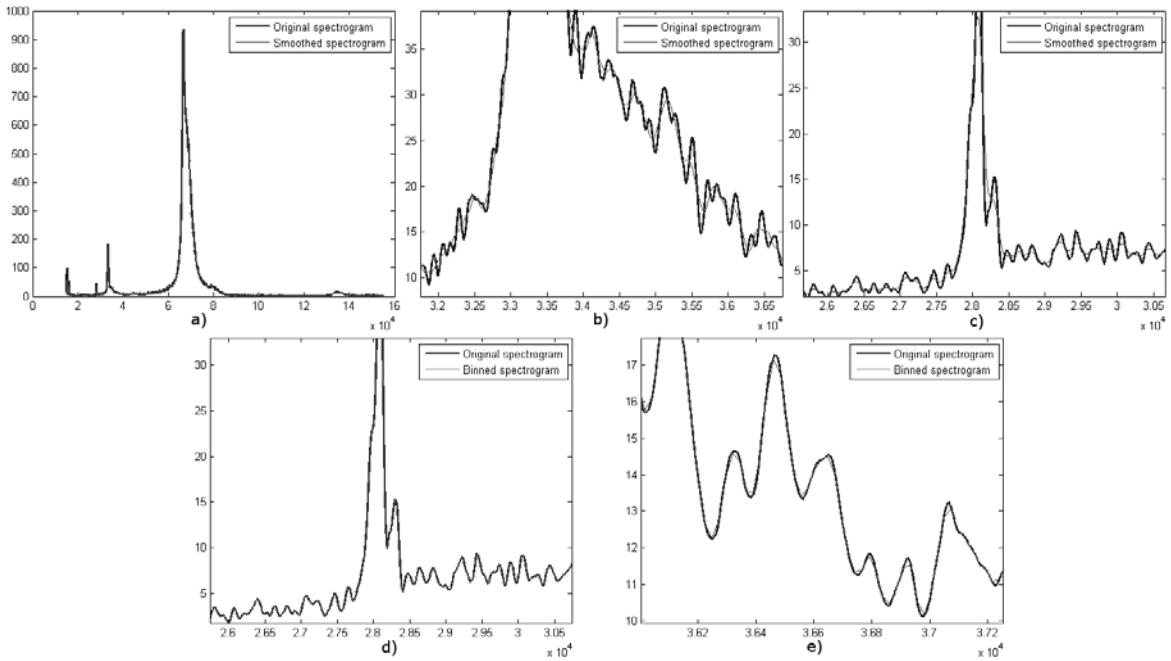

Figure 1 – Examples of usage of denoising and binning methods

The spectrum is being smoothed and range of it depends on chosen mask. Obviously, there are other methods for smoothing and noise reduction. One can use wavelet transformation (for example widely used undecimated discrete wavelet transformation, UDWT), least squares digital polynomial filter (Savitzky and Golay filters) or nonparametric smoothing (locally weighted linear regression with specified window size and type of kernel). Fig. 1 presents examples of smoothing and binning results. Fig. 1a and 1b shows usage of Savitzky and Golay filters smoother (respectively all spectra and only part it it), and Fig. 1c – nonparametric smoother with span of 300 and Gaussian kernel. Fig. 1d and 1e present comparison of binning with, respectively, mask of 4 and 20. One can observe, that in the first case original and binned spectra are almost identical. Usage of binning with mask of 20 gave also good results, however binned spectrum are visibly more angular. It is important to mention, observed original spectrum had 88133 points.

Interpolation process may be defined as unification of measurements points [3] along the given m/z axis, while trimming is cutting of lower or/and upper parts of spectra according to specified boundaries.

Following pre-processing part is baseline removal. It flattens the base profile of a spectrum. It is very important issue, because each mass spectrum contains a base level of intensity (a baseline) which should be identified and subtracted. This process often determines success of further analysis. This kind of algorithms uses such methods and techniques as: simple frame with fixed size and quantilles. An example of baseline process is presented on Fig. 2.

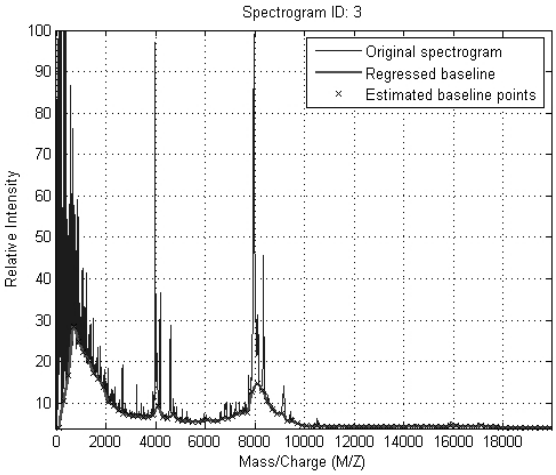

Figure 2 – Example of baseline correction method

Important task in pre-processing is normalization [5] ion intensities to minimize differences between spectra and peaks intensities. The most known methods are: scaling all spectra to total ion current (TIC) value or to constant noise. Another known methods are some mathematical transformations, like ln, log, cube root.

Peaks quantification [1] concerns assessment of signal-to-noise ratio and often involves heights or areas as well.

The last described process is called peaks alignment. This step is required, because none of earlier processes cannot ensure, that founded earlier peaks will match in different samples. This analyze should result in deciding, which peaks in different spectra correspond to the same biological process or molecule.

One of often raised problems is proper sequence establishing. The most popular one ([1] [2] [3] [5]) is: calibration, denoising, baseline subtraction, normalization, peak detection,

quantification and matching. There are, however, works on other methods of mass spectrometry data processing. Randolph et. all [6], for example, discovered method, which do not need of pre-processing at all.

Application of mixture models to detection of peaks of mass spectrometry data

Peak detection is one of the most important tasks in mass spectrometry data processing. Identifying location of peaks is a subject of many research projects. There are few proven methods in that field. Very popular are algorithms based on finding local maxima and minima in denoised spectrum [1]. However, they are not free of disadvantages, like problems with very small peaks. Determining, which of founded peaks are significant may be hard to solve.

Applying of wavelet transformations occurred to be efficient method of peaks detection. The most popular [2] is undecimated discrete wavelet transform (UDWT). It consists in calculation wavelet coefficient including choice of wavelet basis and applying series of linear filters. Afterwards small wavelet coefficient are set to zero and inverse wavelet transform is performed. Type of used wavelet basic is usually inessential [2]. More important is type of thresholding. Soft one [1] is minimizing larger coefficients towards zero, whereas hard one let them unchanged. In mass spectrometry application hard thresholding is more reliable [1], [10]. UDWT may be used in individual spectra (SUDWT) as well as in usage of mean spectra (MUDWT) [2]. Also continuous wavelet transformation (CWT). Usage of CWT do not require sample of number based on the power of two, like in case of DWT. Mexican Hat wavelet is often used in this technique [11].

Dijkstra et. all [9] present another method of mass spectrometry data processing with a basis on a mixture models of log-normal distributions. Similar approach was presented on work of Polanski et. all [3]. This idea seems to be interesting and promising.

Gaussian mixture models fit to needs of mass spectrometry data modeling. However in simple mixture model given sample has usually only one dimension. In case of spectra, samples are two-dimensional (m/z values and intensities). To make usage of EM algorithm, weighted type of algorithm need to be developed.

Values of intensities determine number of repeats of corresponding m/z values. Each single m/z value from x axis (xk) should be repeated yk times to obtain single vector of parameters which can be used in EM algorithm (Eq. (1)).

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} x_n p(k \mid x_n, p_{old})}{\sum_{n=1}^{N} p(k \mid x_n, p_{old})}, k = 1,2,...,K$$

$$(\sigma_k^{new})^2 = \frac{\sum_{n=1}^{N} (x_n - \mu_k^{new})^2 p(k \mid x_n, p_{old})}{\sum_{n=1}^{N} p(k \mid x_n, p_{old})}, k = 1,2,...,K \quad (1)$$

$$\alpha_k^{new} = \frac{\sum_{n=1}^{N} p(k \mid x_n, p^{old})}{N}$$

Obtained results depend on amount of determined components. Calculations were performed for different number of components and values of binning mask. Higher value of binning mask caused matching only the heights peaks. Fig. 3 presents results for binning mask 2 and 4 and respectively 50, 150 and 200 of components.

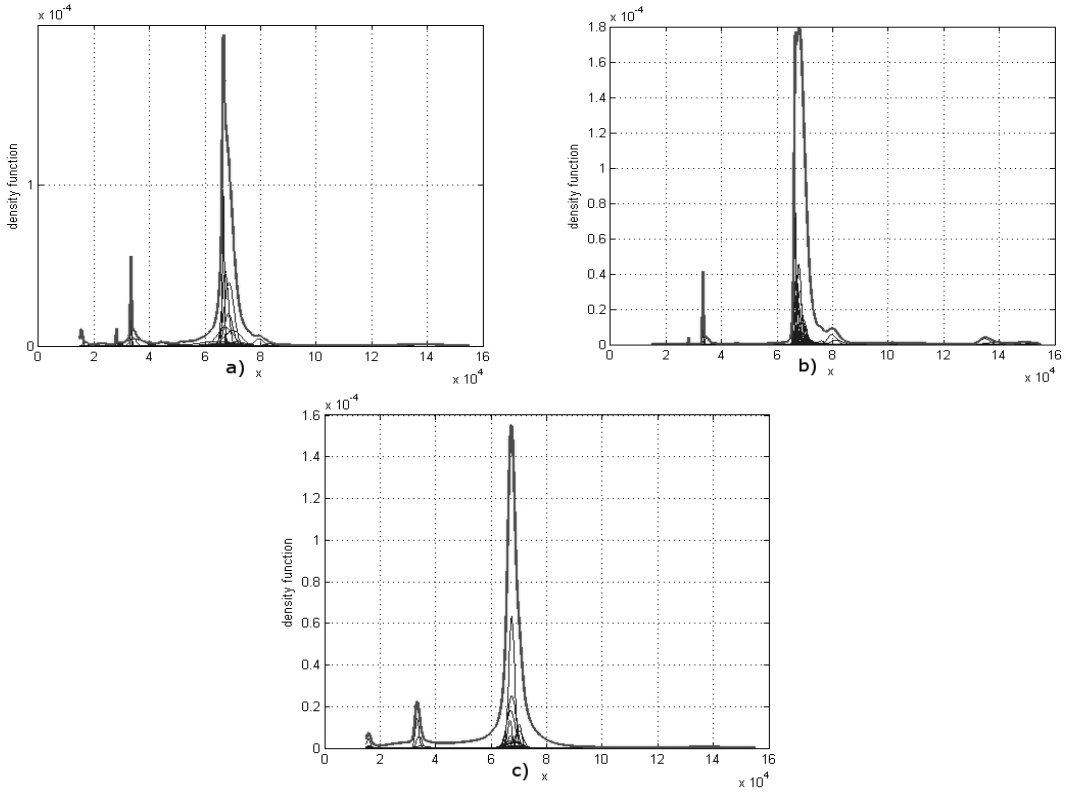

Figure 3 – Usage of EM algorithm for peaks of spectra determination

Important issue is to know the accurate number of components. Bayesian Information Criterion (BIC) is a useful method used for number of components estimation. According to BIC, the optimal number of parameters should maximize the formula Eq.2.

$$-2 \cdot \ln p(x \mid k) \approx BIC = -2 \cdot \ln L + k \ln(n) \tag{3}$$

where

x – the observed sample,

n – the sample size,

k – the number of parameters to estimation

$p(x|k)$ – the likelihood of the observed data given the number of parameters,

L – the maximized value of the likelihood function for the estimated model.

Using the BIC criterion method is time consuming because it needs multiple repetition of EM calculations. Calculations need to be performed several times for all considered number of components. This estimation lasts long but it gives reliable results.

Data classification

Multiple different classifiers might be constructed on the basic of the single learning set. The ideal situation would be to chose the proper classifier on the basic of the number of misclassifications of the new, random observation. However, in reality bad classification probabilities are unknown. They might be estimated from a validation probe. The validation probe is a random sample, independent of the learning probe, where objects' membership to classes are unknown. Misclassification probability of specific classifier is estimated with mistaken classification done by the classifier on the validation probe. Classifier evaluation should be done using observations independent of those from the learning probe. In other cases the classifier is biased.

The ultimate classifier evaluation is done with test probe. It needs to be independent of other probes and it needs to have information about objects' membership to classes. If only one classifier is to be tested or size of the set is small, the validation probe might be omitted. In practice, the usually chosen proportion is the division: 50 % on the learning probe and 25 % each for the validation and test probes [1]. However, the division depends on the specificity of the data set.

One of the most accurate classifiers used for mass spectrometry data is the Support Vectors Machines (SVM). It was proposed by V.N.Vapnik [10,11,12]. The idea of this method is classification with usage of appropriately designated discriminant hyperplane. Searching of such hyperplane needs Mercer theorem and optimization of quadratic objective function with linear restrictions. The SVM idea is to find two parallel hyperplanes, which delimit the wider area do not containing any probe elements. To accept those terms the hyperplanes need to be based on support vectors. If learning sub-sets are not linearly separated, the penalty is introduced. The best separation is obtained for higher dimension space.

The SVM rule takes the form of Eq. 3.

$$f(x) = \mathrm{sgn}(\sum_{\mathrm{sup.}\ vect.} y_i \alpha_i^0 (x_i x) + b^0) \tag{3}$$

where α are Lagrange's coefficients and b is a constant value. For inseparable classes the additional restrictions take the form of Eq. 4.

$$\begin{aligned} x_i w + b &\geq 1 - \xi_i, y_i = 1 \\ x_i w + b &\geq -1 + \xi_i, y_i = -1 \end{aligned} \tag{4}$$

where ξ_i is a constant value $\xi_i \geq 0$

Classifiers used in bioinformatics applications are solved with use of kernel functions. Such construction enables to obtain non-linear shapes of discriminant hyperplanes. One of the most popular kernel function is radial kernel (Eq. 5).

$$K(x_i x') = \exp(-\|x - x'\|^2 / c) \tag{5}$$

Input data-set for classification usually contain several hundreds or even thousands of features. From the statistical point of view using such number of features is unreasonable. There are many reduction and selection techniques available. They attempt to find the smallest data sub-set chosen with defined criteria among the hole data set. Too large number of features has an adverse impact on the classification results. Especially biological data, like mass spectrometry and microarray data fit to this characteristic. Large features number causes increase of computational complexity and lengthen of calculation time. Moreover, large number of features has an influence on low quality of classification. It is due to features correlation. This makes the analysis difficult and the diversification is hard to obtain [2]. Large number of parameters causes also large number of classifier's parameters. It increases its complexity and susceptibility on over learning and decreases its flexibility. The existence of the curse of dimensionality [6] proves, that the complexity of the classifier has an effect on the classification quality. The more complex classifier is, the higher should be the proportion between number of observation and number of features [2].

There are two types of methods:

1. features extraction – data are undergone transformation – new data set is obtained
2. features selection – sub-set of the most optimal data is chosen

One of commonly known features extraction methods is Partial Least Squares (PLS) [15]. It enables also classification. Features selection in PLS method is performed with use of both X and Y data. So it enables using structure of the hole learning data set.

The idea of PLS method is to find latent vectors. Using of latent vectors enables simultaneous analysis and decomposition of X and Y including covariance between X and Y. Such approach makes PLS a special case of Principal Component Analysis (PCA) [13].

Among the most popular features selection method one can find SVM-RFE and traditional T test.

SVM-RFE (Support Vector Machine Recursive Feature Elimination) [14] method is features selection method. Features selection is done with propagation backward method. The procedure starts with full range of input features and features are ranged successively removed. Only one feature is removed in a time. As a rang criterion SVM weights coefficients are used. Therefore SVM-RFE method is closely related to SVM classification.

T test is very common technique of feature selection. The most significant features are chosen according the T test. For each feature a T test range is calculated. The T statistics treats all feature as independent. This assumption is usually not met. However, T test is successfully used for protein data classification.

Biological interpretation

Biological interpretation should be performed with using of proteomic databases. Results achieved from classification is transferred to the application. Integration with several big biological databases makes results of the application always up-to-data.

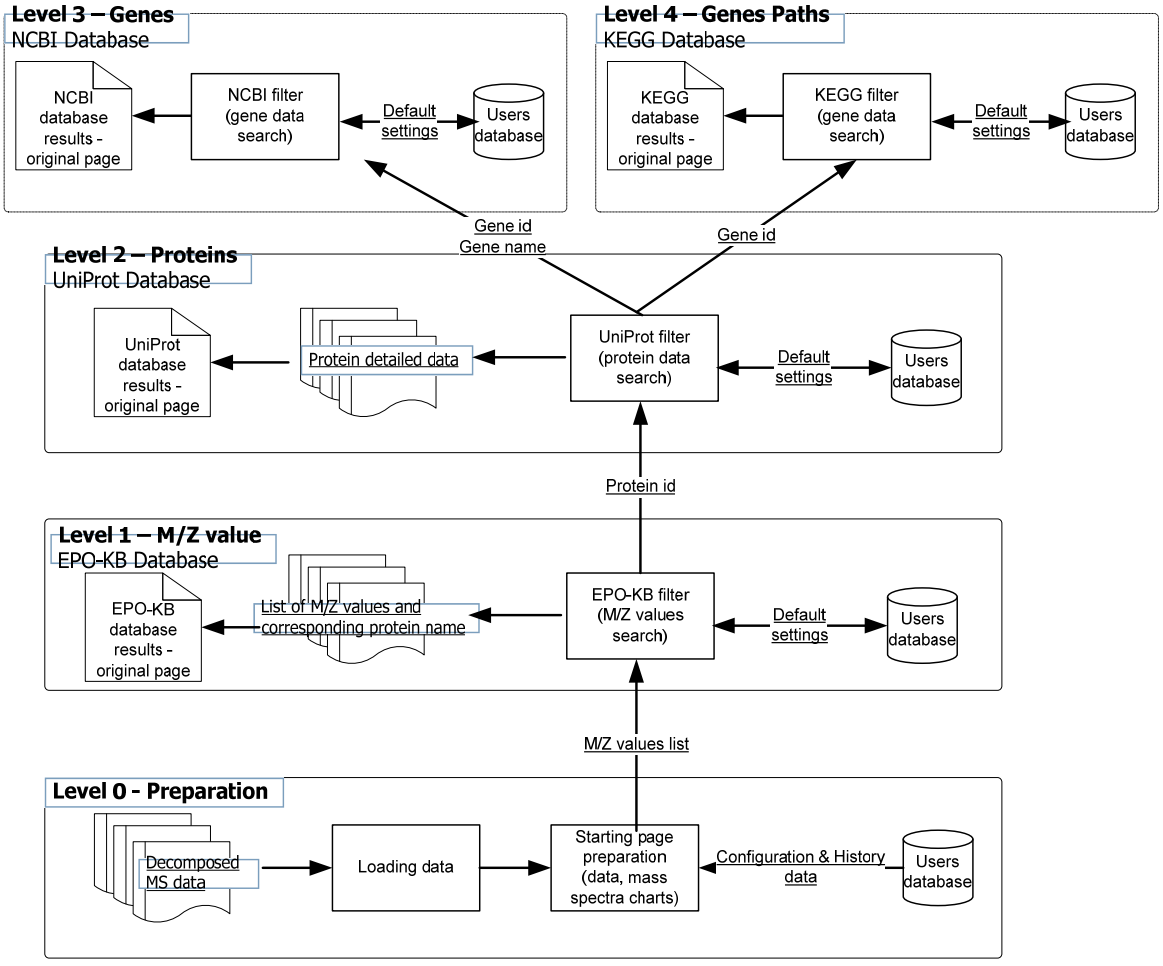

Figure 4 – Usage of EM algorithm for peaks of spectra determination

The example path of biological analysis application is presented at Fig.4. It has been divided into four steps, each of them is responsible for different level of biological context. Levels should be achieved sequentially.

At level0 user is able to load data and give detailed search criteria. Those criteria includes: accuracy, species, the MS platform, the possibility of double and triple charges. Searching is based on M/Z values, which are transferred from classification module.

Level1 is based on EPO-KB (Empirical Proteomic Ontology Knowledge Base) database [14], [15]. Names of proteins and peptides are found on the basis of given M/Z values with a specified percentage tolerance. User can also see the original results in the EPO-KB service.

Level2 is a protein level and data presented here are obtained from an UniProt [16,17] database. Displayed results contains detailed information about proteins, such as entry name, status of reviewing process, organism, gene names and identifiers, features or GO annotations. It is also possible to see the original results returned by the database.

Level3 is a genes level and it gives information about genes coding a particular protein chosen at an previous level2. Presented data are based on NCBI service [9]. Searching is based on the gene identifier and it returns precise information about a particular gene, its role, status, lineage and related data. Level4 is based on gene pathways data. It is integrated with the KEGG database (Kyoto Encyclopedia of Genes and Genomes) [18]. Level4 gives details about genes pathways, structures, sequences, references to other databases.

The results of biological analysis are presented in Fig 5.

Wartość MZ	Nazwa proteiny
9400.481	• (42.5) apolipoprotein c-iii
9637.234	• Nie znaleziono protein spełniających zadane kryteria
9282.325	• (9.7) platelet basic protein • (31.7) c-c motif chemokine 13
9239.571	• (48.1) haptoglobin • (52.4) platelet basic protein
9375.544	• Nie znaleziono protein spełniających zadane kryteria
9718.618	• Nie znaleziono protein spełniających zadane kryteria
9161.69	• (29.8) haptoglobin
8925.394	• (5.4) complement c3 frag • (9.6) complement c3 • (20.6) apolipoprotein a-ii • (25.4) vitronectin frag
9383.259	• Nie znaleziono protein spełniających zadane kryteria
9415.935	• (27.1) apolipoprotein c-iii
9465.652	• (22.7) apolipoprotein c-iii
9126.531	• (3.5) haptoglobin
9525.366	• Nie znaleziono protein spełniających zadane kryteria

Figure 5 – Usage of EM algorithm for peaks of spectra determination

Summary

Presented analyses indicate several aspects of mass spectra processing. Analysis of several different data sets proves the ability of the tool to adjust parameters sets to data. Using Gaussian Mixture Models and EM algorithm enables analysis of different types of mass spectra data sets and considers interactions between overlapped peaks.

Proposed paths of analysis indicate, that analysis of mass spectrometry data should be complex and it should cover several different aspects, like preprocessing, peaks detection, classification and biological interpretation.

Biological databases are priceless tool which enable up-to-date analysis of the actual composition of analyzed samples.

Bibliography

1. Coombes K. Pre-processing mass spectrometry data / K. Coombes, K. Baggerly, J. Morris // Fundamentals of Data Mining in Genomics and Proteomics / W Dubitzky, M Granzow and D Berrar [eds.]. – Boston : Kluwer,. 2007. – P. 79-99.
2. Feature extraction and quantification for mass spectrometry data in biomedical applications using the mean spectrum / Morris J., Coombes K., Kooman J. [eds.] // Bioinformatics. – 2005. – № 21(9). – P. 1764-1775.
3. Application of the Gaussian mixture model to proteomic MALDI-ToF mass spectra / Polanski A., Polanska J., Pietrowska M. [eds.] // Journal of Computational Biology. – 2007, Gliwice.
4. Denoier R. Computational Genome Analysis. An introduction / R. Denoier, S. Tavare, M. Waterman. – USA : Springer. 2005.
5. Processing MALDI mass spectra to improve mass spectral direct tissue analysis / Norris J., Cornett D., Mobley J. [eds.]. – USA : National institutes of health, 2007.
6. Quantifying peptide signal in MALDI-TOF mass spectrometry data / Randolph T., Mithcell B., McLerran D., [eds.] // Molecular & Cellular Proteomics 4:1990-1999. The American Society for Biochemistry and Molecular Biology. –2005.
7. Polański A. Bioinformatics / A. Polański, M. Kimmel. – Berlin Heidelberg.F : Springer, 2007.
8. Plechawska M. Comparing and similarity determining of Gaussian distributions mixtures. Materials of SMI Conference / M. Plechawska. – Swinoujscie,2008.
9. Peak quantification in surface-enhanced laser desorption/ionization by using mixture models / Dijkstra M., Roelofsen H., Vonk R., Jansen R. // Proteomics. – 2006. – 6. – P. 5106-5116.
10. Vapnik V. A training algorithm for optimal margin classifiers. Fifth Annual Workshop on Computational Learning Theory / V. Vapnik, B. Boster, I.. Guyon. – 1992. – P. 114-152.
11. Vapnik V.N. The Nature of Statistical Learning Theory / Vapnik V.N. – Springer, 1995.
12. Vapnik V.N. Statistical Learning Theory / Vapnik V.N. – Wiley, 1998.
13. Mao J. Statistical pattern recognition: a review / J. Mao, A.K. Jain, R.P.W. Duin. // IEEE Trans. PAMI. – 22(1). – 2000. – P. 4-37.
14. Wold H. Estimation of principal components and related models by iterative least squares. Multivariate Analysis / H. Wold. – New York : Academic Press, 1996. – P. 391-420.
15. Barnhill S. Gene selection for cancer classification using support vector machines / S. Barnhill, V. Vapnik, I. Guyon, J. Weston. – Machine Learning, 2002. – P. 389-422.
16. Zhang S.Q. Peak detection with chemical noise removal using Short-Time FFT for a kind of MALDI Data. Proceedings of OSB 2007 / S.Q. Zhang // Lecture Notes in Operations Research. – 2007. – № 7. – P. 222-231.
17. Randolph T. Quantifying peptide signal in MALDI-TOF mass spectrometry data / T. Randolph // Molecular & cellular proteomics : MCP. – 4(12):1990-9 2005.
18. Improved method for peak picking in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry / Kempka M., Sjodahl J., Bjork A., Roeraade J. // Rapid Commun. Mass Spectrom. – 2004. – Vol. 18. – P. 1208-1212.

М. Плехавска-Войчик

Комплексные методы спектрального анализа MALDI-TOF данных

В статье представлен обзор методов MALDI-TOF данных с многошаговым спектрометрическим анализом. Эта сложная задача должна охватывать несколько платформ. Чтобы получить полезные результаты, важно сделать всеобъемлющий анализ. Подобный анализ должен охватывать анализ предварительной обработки сигналов и анализ баз данных и поиска пиков классификации.

Статья поступила в редакцию 31.05.2011.