STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

Małgorzata PLECHAWSKA-WÓJCIK
Politechnika Lubelska, Instytut Informatyki

BIOLOGICAL INTERPRETATION OF THE MOST INFORMATIVE PEAKS IN THE TASK OF MASS SPECTROMETRY DATA CLASSIFICATION

Summary. The article presents results of integration of classification of MALDI--ToF mass spectrometry data with proteomic databases. This biological interpretation of classification results is based on popular biological databases, such as EPO-KB, UniProt, NCBI. The classification is performed with Support Vector Machines and dimension reduction techniques.

Keywords: SVM, MALDI-ToF, GMM

INTERPRETACJA BIOLOGICZNA NAJBARDZEJ INFORMACYJNYCH PIKÓW W KLASYFIKACJI DANYCH SPEKTROMETRYCZNYCH

Streszczenie. Artykuł przedstawia integrację klasyfikacji danych spektrometrycznych typu MALDI-ToF z białkowymi bazami danych. Ta biologiczna interpretacja wyników klasyfikacji oparta jest na popularnych biologicznie bazach danych, takich jak: EPO-KB, UniProt, NCBI. Klasyfikacja została przeprowadzona z wykorzystaniem Maszyny Wektorów Wspierających (SVM) oraz metod redukcji wymiarowości.

Słowa kluczowe: SVM, MALDI-ToF, mieszaniny rozkładów Gaussa

1. Introduction

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.

The aim of the work is to present a tool dedicated to comprehensive mass spectra analysis. A data set of cancer patients was proceeded and presented results show steps of the analysis. The idea of the work is to prove, that it is possible to create the tool, which is able to support biologists in data analysis, concerning creating and solving the mass spectrometry data models, choosing the most informative peaks and giving its biological interpretation.

There are many tools and applications designed to support spectra analysis. However, most of them are concentrated on mathematical analysis of the signal. This process is complex and it consists of pre-processing [10, 20] (denoising, baseline correction, normalization), peaks detection and alignment. Most of the tools realize those functions. They use different methods. Denoising can be done with local maxima smoothing (Cromwell package [20], PROcess [21], ProteinChip Software [22]), removing some detected peaks (LIMPIC [23]) or moving average with dedicated filters (OpenMS [24]). The core of the analysis - peaks detection, can be obtained on the basis of local maxima and signal to noise ratio ([22]), area under the peaks curve ([21]), the height (SpecAlign [25]) and the shape of peaks (OpenMS [26]). Most of mass spectra analysis tools concentrate only on solving mathematical models. The authorial tool, which gave results presented in this article is more comprehensive. Results of peaks detection, obtained with using Gaussian mixture model analyzing, are subjected to further operations: classification and biological interpretation.

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 learn-

ing set is composed of the object representation and the class label. This object representation is vector of observation features. The whole set is divided into c separated subsets. One subset observations are numbered among one of c classes. Such supervised learning is widely used in biomedical applications.

2. Prediction models

The essential task is to construct the classifier on the basis of the data set. It is possible to construct multiple different classifiers on the basis of the same data. The point is to find the most suitable one. It would be ideal to chose the best classifier on the basic of its ability to classify new observations. However, such probabilities are unknown. The most obvious way to solve it is to divide whole data set into training and validation probes. The validation probe is a random sample, independent of the learning probe. It is used to assume misclassification probabilities of specified classifiers. The most important is to keep the validation probe observations independent of those from the learning probe. In other cases the classifier is biased and it might give results oriented to the data from the particular data set. Such classifier will give good results for the data from the validation probe. However, it could be give poor result for any other data from other data sets. The ultimate classifier evaluation might be done with additional, test probe. It needs to be independent of other probes and 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.

The most popular classification quality measures are:

- classification accuracy (a proportion of correctly classified sets),
- error rate (a proportion of misclassified sets),
- TP (True Positives) – the number of correctly classified positive sets,
- TN (True Negatives) – the number of correctly classified negative sets,
- FP (False Positives) – the number of incorrectly classified positive sets,
- FN (False Negatives) – the number of incorrectly classified negative sets,
- sensitivity (eq. 1) (the classifier ability to identify the phenomenon where it really exists),
- specificity (eq. 2) (the ability to reject truly false results - opposed to sensitivity),
- ROC (a chart of dependency between values: 1- specificity and the sensitivity) and AUC (the area under the curve).

$$sensitivity = \frac{TP}{FN + TP} \tag{1}$$

$$specificity = \frac{TN}{TN + FP} \tag{2}$$

3. Characteristic of the data set

The analyzed dataset is composed of Maldi-Tof (Matrix-Assisted Laser Desorption Ionization - Time Of Flight) mass spectra files. Serum samples were collected between V.2006 to I.2008 by Department of Experimental and Clinical Radiobiology in Comprehensive Cancer Centre Maria Sklodowska-Curie Memorial Institute. Samples were taken from patients with breast cancer diagnosis and from healthy donors. In the study there were 92 patients with I and II rate of disease progression. The average age of patients was 58.5 (patients were 31-74 years old). The control group was composed of 104 healthy women in good general health state. The average age in this group was 54 (women were 32-77 years old).

Samples were collected twice from each patient. Each of those two obtained samples was analyzed in mass spectrometry two times. As a result for every patient four data files were obtained. In this way four spectra are generated per each patient. Such technique enables noise reduction and data quality increase.

Each of the data set file contains 45 thousands of points. Typical mass spectrum is composed of two data vectors: M/Z value (X axis) and intensities (Y axis). The aim of the analysis is to detect peaks and find its biological interpretation using biological databases.

4. Preparation of the data model

Spectra analysis is composed of several steps. All operations need to be processed carefully, because improperly processed data have strong negative influence on further steps of analysis.

The first step is outliers removing. It is performed among four spectra of one person. After outliers removing mean spectra are calculated for each person separately.

The next group of operations is preprocessing. Preprocessing steps involve: binning, interpolation, normalization, baseline correction, normalization, denoising, peaks detection [10] and alignment [9]. One of the most important preprocessing steps is denosing, especially baseline correction. Baseline is a special case of noise, intensifying especially in initial part of the spectrum, where M/Z values are low. Baseline correction flattens and averages the spec-

trum. It is usually performed with multiple shifted windows with defined width. Normalization and interpolation are techniques used in analyzing and comparing several spectra simultaneously. Interpolation is the technique of measurements points unification [12]. This unification is performed along m/z axis of all spectra. Normalization [11,13] is scaling all spectra to a single value of area under the curve or total ion current (TIC). An example of analyzed spectrum with baseline correction result is presented at fig. 1.

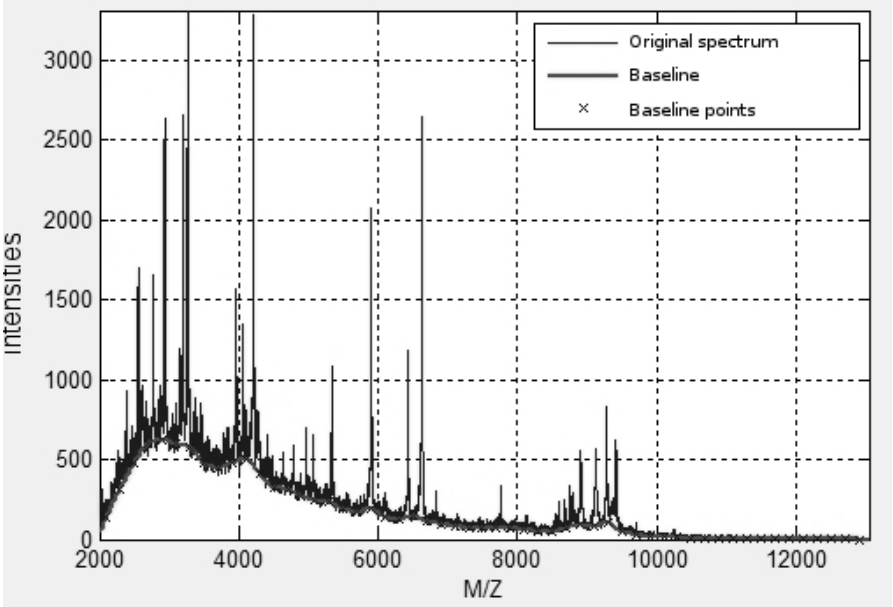

Fig. 1. Preprocessing of mass spectra
Rys. 1. Wstępna analiza widm

Preliminary dimension reduction might be performed by peaks detection. After preprocessing the mean spectrum is calculated. It is modeled with Gaussian mixture model (GMM). Number of components was estimated with Bayesian Information Criterion (BIC) and it was set on 200 components.

The fitting is done with Expectation-Maximization algorithm (EM) performing maximizing the likelihood function. A typical mixture model is a combination of a finite number of probability distributions (eq. 3)

$$f^{mix}(x,\alpha_1,\ldots,\alpha_K,p_1,\ldots,p_K)=\sum_{k=1}^{K}\alpha_k f_k(x,p_k) \tag{3}$$

where K is the number of components in the mixture and $\alpha_k, k=1,2,..K$ are weights of particular component, $\sum_{k=1}^{K}\alpha_k=1$. Gaussian distribution (eq. 4) is given with two parameters: mean μ_k and standard deviation σ_k. Distributions in the mixture are also specified with additional parameters – weights, which determine their contribution to the whole mixture.

$$f_k(x_n,\mu_k,\sigma_k)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_n-\mu_k)^2}{2\sigma_k^2}\right] \tag{4}$$

The Expectation-Maximization (EM) algorithm is nonlinear method and is composed of two main steps performed in the loop. The expectation step (E) consists in calculation of distribution of hidden variables (eq. 5)

$$p(k \mid x_n, p^{old}) = \frac{\alpha_k^{old} f_k(x_n, p^{old})}{\sum_{\kappa=1}^{K} \alpha_k^{old} f_k(x_n, p^{old})} \tag{5}$$

The maximization step (M) calculates new mixture parameters values. It is given with (eq. 6).

$$\begin{aligned} \mu_k^{new} &= \frac{\sum_{n=1}^{N} x_n p(k \mid x_n, p_{old})}{\sum_{n=1}^{N} p(k \mid x_n, p_{old})}, k = 1,2,\ldots,K \\ (\sigma_k^{new})^2 &= \frac{\sum_{n=1}^{N} (x_n - \mu_k^{new})^2 p(k \mid x_n, p_{old})}{\sum_{n=1}^{N} p(k \mid x_n, p_{old})}, k = 1,2,\ldots,K \\ \alpha_k^{new} &= \frac{\sum_{n=1}^{N} p(k \mid x_n, p^{old})}{N} \end{aligned} \tag{6}$$

The decomposition results are used as a Gaussian mask which was put on every single spectrum in the data set. This gives new values consisting the spectra. Dimensions of spectrometry data decreased to the value of GMM components number. The result matrix obtained after those steps was: $n \times k$, where n denoted number of spectra and k - number of components.

The resultant matrix was the input data to the further dimension reduction and classification.

5. SVM classifier and dimension reduction techniques

The Support Vectors Machines (SVM) is young, but widely used classifier. It was proposed by V.N.Vapnik [3,4,5]. The idea of this method is classification with usage of appropriately designated discriminant hyperplane. If learning sub-sets are fully separable, 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 some of the probe elements. Such elements are called support vectors. The discriminant hyperplane is put in the middle of the resultant area. 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. 7).

$$f(x) = \operatorname{sgn}\Big(\sum_{\text{sup. } vect.} y_i \alpha_i^0 (x_i x) + b^0 \Big) \tag{7}$$

where α are Lagrange's coefficients and b is a constant value. For inseparable classes the additional restrictions take the form of (eq. 8).

$$\begin{aligned} x_i w + b &\geq 1 - \xi_i, y_i = 1 \\ x_i w + b &\geq -1 + \xi_i, y_i = -1 \end{aligned} \tag{8}$$

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

The more complicated classification problems are solved with use of kernel functions. Such construction enables to obtain non-linear shapes of discriminant hyperplanes. The SVM rule with kernel takes the form of (eq. 9).

$$f(x) = \text{sgn}(\sum_{\text{sup. } vect.} y_i \alpha_i^0 K(x_i x) + b^0) \tag{9}$$

where $K(x_i x)$ is a kernel. One of the most popular kernel function is radial kernel (eq. 10).

$$K(x_i x') = \exp(-\|x - x'\|^2 / c) \tag{10}$$

If data-set contains several hundreds or even thousands of features, it is unable to gain proper classification results. In such case reduction or selection techniques should be used. They attempt to find the smallest data sub-set chosen with defined criteria among the whole 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 [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 [7].

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 and classification methods is Partial Least Squares (PLS) [7]. PLS features selection is performed with use of both X and Y data. So it enables using structure of the whole learning data set. The idea of PLS 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) [5].

The decomposition of X and Y is done to low-dimensional space of hidden variables. Independent variables X are decomposed according (eq. 11).

$$X = TP^T + E_x \tag{11}$$

where $T^T T = I$, I - identity matrix, T - score matrix and P - loading matrix. A product of T and P gives good estimation of X matrix.

Dependent variables Y are decomposed as(eq. 12).

$$Y = UQ^T + E_y \tag{12}$$

The final model of PLS describing $Y \Leftrightarrow X$ regression is (eq. 13).

$$Y = X(PB_1Q^T) + E = XB + E \tag{13}$$

SVM-RFE (Support Vector Machine - Recursive Feature Elimination) [8] 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.

In SVM-RFE procedure SVM classification might be formulated as in (eq. 14).

$$\begin{aligned} &\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i^2 \\ &y_i(wz_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \end{aligned} \tag{14}$$

Eq. 14 is solved with (eq. 15).

$$\min_{\alpha_i} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \tilde{k}(x_i x_j) - \sum_{i=1}^{n}\alpha_i \tag{15}$$

where $\tilde{k}(x_i x_j)$ is a kernel function.

The SVM-RFE objective function is (eq. 16).

$$J = \frac{1}{2}\|w\|^2 \tag{16}$$

Changes in the objective function caused by features elimination may be written using the Taylor series (eq. 17). $\Delta J(i)$ in the optimal point takes the value $\Delta J(i) = (\Delta w_i)^2$, where w_i is the i[th] removed feature.

$$\Delta J(i) = \frac{\partial J}{\partial w_i}\Delta w_i + \frac{\partial^2 J}{\partial w^2{}_i}(\Delta w_i)^2 \tag{17}$$

Very common technique of feature selection is T test. The most significant features according the T test are chosen. For each feature a T test range is calculated with (eg. 18).

$$c_i = \frac{\left|\mu_i^+ - \mu_i^-\right|}{\sqrt{\frac{(\sigma_i^+)^2}{n^+} + \frac{(\sigma_i^-)^2}{n^-}}} \tag{18}$$

where μ_i^+, μ_i^- denote the mean values for i[th] feature calculated for respectively positive and negative samples. Similarly σ_i^+, σ_i^- denote standard deviations and n^+, n^- denote numbers of positives and negatives learning samples.

The T statistics treats all feature as independent. This assumption is usually not met. However, T test is successfully used for protein data classification.

6. Selection of classifier parameters

Classification and feature reduction or selection methods has parameters, which need to be set before the classifier construction. The classification analysis was performed using all three presented dimension reduction techniques. The classification was done with SVM classifier – both linear and with radial kernel. Three dimension reduction techniques were also used (PLS, SVM-RFE, T test). SVM parameters which need to be estimated are: value of box constraints (C) for the soft margin and the scaling factor (sigma) for radial kernel. Important issue is also to find the most accurate number of features. To find the most accurate values, division of the data set into testing and learning subsets and classification calculations need to be repeated several hundred times. All calculations were done in Matlab environment.

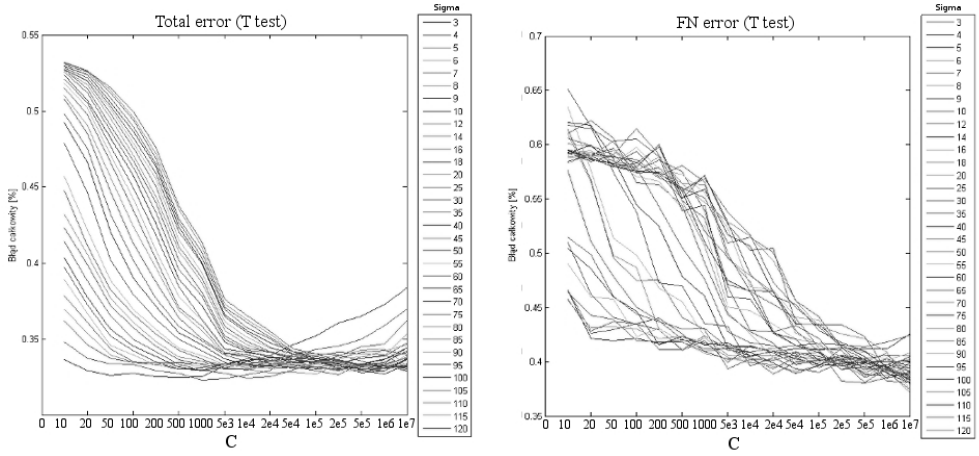

Fig. 2. Estimation of C parameter
Rys. 2. Oszacowanie parametru C

According simulation results SVM parameters remain similar for all tested dimension reduction methods. Simulation study was performed to calculate not only the total error but also FN and FP values. Fig. 2 presents results for C estimation and fig. 3 – for sigma estimation. On both figures the middle line is the obtained ratio and the upper and lower denotes the confidence interval. In further calculations C value was set to 1e6 and sigma – to 12. Figure 4 presents case study for features number determination. As results show, classification errors

is descending until they gain optimal value. After that errors are increasing what suggest overloading. The optimal value was set to 8 features.

Important issue was also to detect and remove features which have the highest dependency rate. Those features raised in the decomposition process. Such features indicate peaks which were detected with too high nearness. Such features are removed or selected to make classification process more reliable.

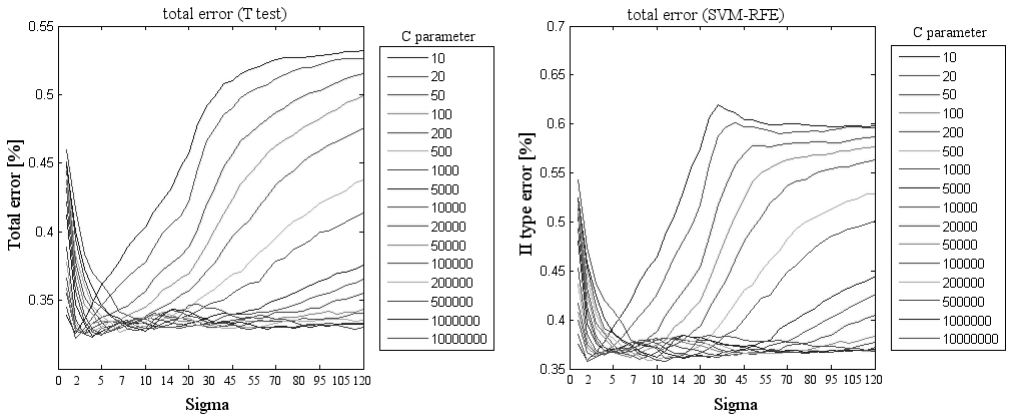

Fig. 3. Estimation of Sigma parameter
Rys. 3. Oszacowanie parametru Sigma

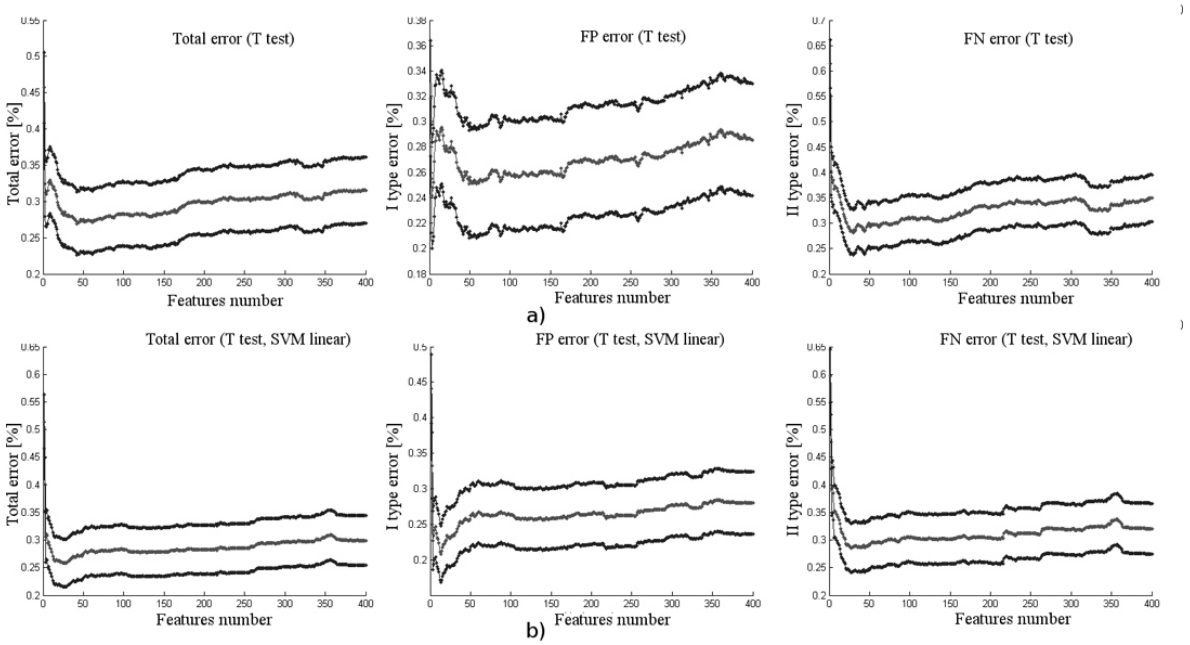

Fig. 4. Estimation of number of features
Rys. 4. Oszacowanie liczby cech

Results of simulation gave parameters, which were used in further analysis. The ROC curve (fig. 5a) shows the final result of classification. Similar simulation for the same data set was performed with using other, popular peak detection method based on local maxima. Those results occurred less efficient (fig. 5b).

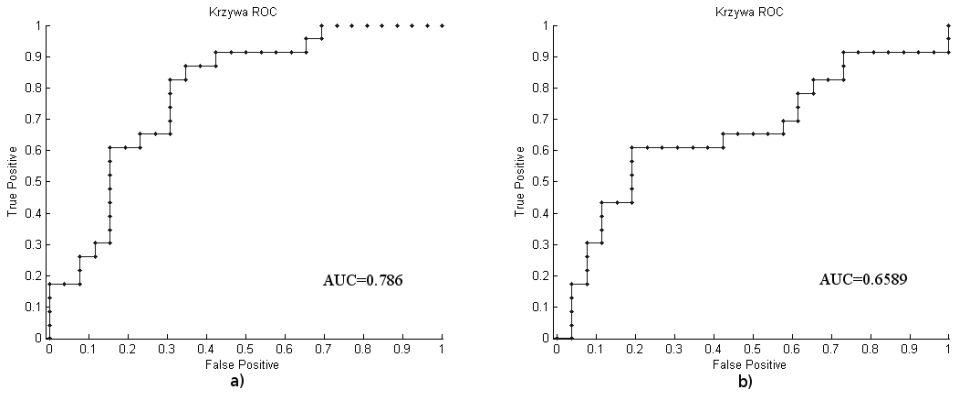

Fig. 5. The ROC curve
Rys. 5. Krzywa ROC

7. Biological interpretation

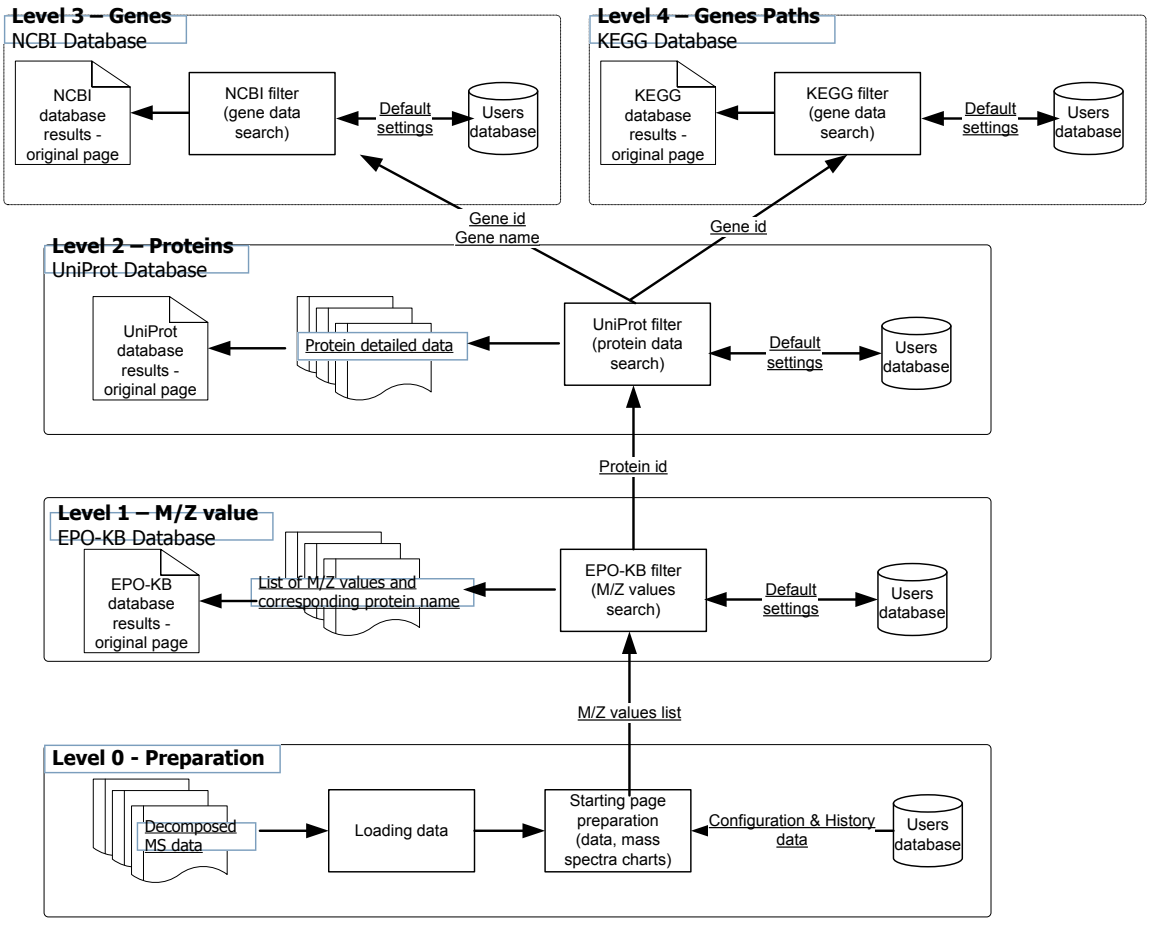

Fig. 6. Flowchart of the biological context module
Rys. 6. Schemat działania modułu informacji biologicznej

Biological interpretation is done with application based on proteomic databases. Results achieved from classification is transferred to the application. Integration with four big biological databases makes results of the application always up-to-data.

The application has been divided into four steps, each of them is responsible for different level of biological context (fig. 6). Levels need to be achieved sequentially.

At level 0 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.

Level 1 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.

Level 2 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.

Level 3 is a genes level and it gives information about genes coding a particular protein chosen at an previous level 2. Presented data are based on NCBI service [18]. Searching is based on the gene identifier and it returns precise information about a particular gene, its role, status, lineage and related data. Level 4 is based on gene pathways data. It is integrated with the KEGG database (Kyoto Encyclopedia of Genes and Genomes) [19]. Level 4 gives details about genes pathways, structures, sequences, references to other databases.

The results of biological analysis are presented at fig. 7 and fig. 8.

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

Fig. 7. List of proteins
Rys. 7. Lista białek

Obtained results, according a professional biologist, are reasonable. Missing data (fig. 6) indicates false discovery errors. Such errors appears in classification task and the only way to prevent them is to perform checking on different level. A good option is to perform automatic biological interpretation. Using EPO-KB database which is serum proteins database highly reduces possibility of unreasonable results obtaining. This method gives biological interpretation of results and it is one of the possible solutions. The other one is to experimentally check the biological composition.

The efficiency of the method was also checked with different data sets obtained from other experiments published by other scientists. Results [27] gives high rate of repeatability and correspondence.

Accesion	EntryName	Status	ProteinNames	GeneNames	Organism	Length	Szczegóły
P02656	APOC3_HUMAN	reviewed	Apolipoprotein C-III (Apo-CIII) (ApoC-III) (Apolipoprotein C3)	APOC3	Homo sapiens (Human)	99	Więcej
P02647	APOA1_HUMAN	reviewed	Apolipoprotein A-I (Apo-AI) (ApoA-I) (Apolipoprotein A1) [Cleaved into: Apolipoprotein A-I(1-242)]	APOA1	Homo sapiens (Human)	267	Więcej
P06727	APOA4_HUMAN	reviewed	Apolipoprotein A-IV (Apo-AIV) (ApoA-IV) (Apolipoprotein A4)	APOA4	Homo sapiens (Human)	396	Więcej
A3KPE2	A3KPE2_HUMAN	unreviewed	Apolipoprotein C-III (Apolipoprotein C-III variant 2) (Apolipoprotein C-III variant 3) (Apolipoprotein C-III, isoform CRA_a)	APOC3 hCG_41334	Homo sapiens (Human)	99	Więcej
Q6Q788	APOA5_HUMAN	reviewed	Apolipoprotein A-V (Apo-AV) (ApoA-V) (Apolipoprotein A5) (Regeneration-associated protein 3)	APOA5 RAP3 UNQ411/PRO773	Homo sapiens (Human)	366	Więcej
B0YIW2	B0YIW2_HUMAN	unreviewed	Apolipoprotein C-III variant 1	APOC3	Homo sapiens (Human)	117	Więcej

Fig. 8. Proteins level
Rys. 8. Poziom białek

8. Summary

Mass spectrometry data need for special processing and analyzing. Data specificity makes it hard to analyze and classify. Proper classification techniques need to be use and special dimension reduction methods should be employed. For each of them parameters needs to be estimated. SVM classifier and its variants is nowadays one of the most popular classification technique among proteomic research.

Using application integrating biological databases it is possible to gain information about biological context of analyzed data set. It enables collecting essential information in one place.

BIBLIOGRAPHY

1. Cwik J., Koronacki J.: Statystyczne systemy uczące się. Akademicka Oficyna Wydawnicza Exit, Warszawa 2008, s. 239÷245.
2. Stąpor K.: Automatyczna klasyfikacja obiektów. Akademicka Oficyna Wydawnicza Exit, Warszawa 2005, s. 35÷52.
3. Vapnik V., Boster B., Guyon I.: A training algorithm for optimal margin classifiers. Fifth Annual Workshop on Computational Learning Theory, 1992, p. 114÷152.
4. Vapnik V.N.: The Nature of Statistical Learning Theory. Springer, 1995.
5. Vapnik V.N.: Statistical Learning Theory. Wiley, 1998.
6. Mao J., Jain A.K., Duin R.P.W.: Statistical pattern recognition: a review. IEEE Trans. PAMI, 22(1), 2000, p. 4÷37.
7. Wold H.: Estimation of principal components and related models by iterative least squares. Multivariate Analysis, New York: Academic Press 1996, p. 391÷420.
8. Barnhill S., Vapnik V., Guyon I., Weston J.: Gene selection for cancer classification using support vector machines. Machine Learning, 2002, p. 389÷422.
9. Morris J., Coombes K., Kooman J., Baggerly K., Kobayashi R.: Feature extraction and quantification for mass spectrometry data in biomedical applications using the mean spectrum. Bioinformatics, 21(9), 2005, p. 1764÷1775.
10. Norris J., Cornett D., Mobley J., Anderson M., Seeley E., Chaurand P, Caprioli R.: Processing MALDI mass spectra to improve mass spectral direct tissue analysis. National institutes of health, 260(2-3), USA 2007, p. 212÷221.
11. Karpievitch Y.V., Hill E.G., Smołka A.J., Morris J.S., Coombes K.R., Baggerly K.A., Almeida J.S.: PrepMS: TOF MS Data Graphical Preprocessing Tool. Bioinformatics, 23, 2007, p. 264÷265.
12. Plechawska M.: Simultaneous analysis of multiple Maldi-TOF proteomic spectra using the mean spectra. SMI 2009. Polish Journal of Environmental Studies. wyd. Hard Olsztyn. Vol. 18, No. 3B, 2009, p. 1230÷1244.
13. Hilario M., Kalousis A., Pellegrini C., Müller M.: Processing and classification of protein mass spectra. Mass Spectrom Rev., 25, 2006, p. 409÷449.
14. Lustgarten, J.L., et al.: EPO-KB: a searchable knowledge base of biomarker to protein links. Bioinformatics, 24(11), 2008, p. 1418÷1419.
15. Lustgarten, J.L., et al.: Knowledge-based variable selection for learning rules from proteomic data. Bioinformatics 10(Suppl 9): S16, 2009.
16. UniProt Consortium: The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res., 2010, p. D142÷D148.

17. Jain E., Bairoch A., Duvaud S., Phan I., Redaschi N., Suzek B. E., Martin M. J., McGarvey P., Gasteiger E.: Infrastructure for the life sciences: design and implementation of the UniProt website.BMC Bioinformatics 10:136, 2009.
18. Wheeler D. L. et al.: Database resources of the National Center for Biotechnology Information. Nuc. Acids Res. 37, 2009, p. D5÷D15.
19. Kanehisa M., Araki M., Goto S., Hattori M., Hirakawa M., Itoh M., Katayama T., Kawashima S., Okuda S., Tokimatsu T., Yamanishi Y.: KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, 2008, p. D480÷D484.
20. Plechawska M., Polańska J., Polański A., Pietrowska M., Tarnawski R., Widlak P., Stobiecki M., Marczak Ł.: Analyze of MALDITOF proteomic spectra with usage of mixture of Gaussian distributions. Cyran K., Kozielski S., Peters J., Stańczyk U., Wakulicz-Deja A. (eds.): Man-machine interactions, Advances in Intelligent and Soft Computing. Springer, Berlin 2009, p. 113÷120.
21. Li X., Gentleman R., Lu X., Shi Q., Iglehart J. D., Harris L., Miron A: SELDI-TOF mass spectrometry protein data, Statistics for Biology and Health, Part I, Bioinformatics and computational biology solutions using R and Bioconductor, 2005, p. 91÷109.
22. Ciphergen Biosystems: ProteinChip Software Operation Manual, Fremont, CA 94555, 2002.
23. Mantini D., Petrucci F., Pieragostino D., Del Boccio P., Di Nicola M., Di Ilio C., Federici G., Sacchetta P., Comani S., Urbani A.: LIMPIC: a computational method for the separation of protein signals from noise, BMC Bioinformatics, 8, 101, 2007.
24. Sturm M., Bertsch A., Gropl C., Hildebrandt A., Hussong R., Lange E., Pfeifer N., Schulz-Trieglaff O., Zerck A., Reinert K., Kohlbacher O.: OpenMS – An open-source software framework for mass spectrometry, Bioinformatics, 9, 163, 2008.
25. Wong J. W., Cagney G., Cartwright H. M.: SpecAlign - processing and alignment of mass spectra datasets, Bioinformatics, 21, 2005, p. 2088÷2090.
26. Sturm M., Bertsch A., Gropl C., Hildebrandt A., Hussong R., Lange E., Pfeifer N., Schulz-Trieglaff O., Zerck A., Reinert K., Kohlbacher O.: OpenMS – An open-source software framework for mass spectrometry, Bioinformatics, 9, 163, 2008.
27. Plechawska M.: Simultaneous analysis of multiple Maldi-TOF proteomic spectra using the mean spectra. Comprehensive analysis of mass spectrometry data – a case study. SMI 2011. Article under construction.

Recenzenci: Dr inż. Michał Kawulok
Prof. dr hab. inż. Andrzej Polański

Wpłynęło do Redakcji 31 stycznia 2011 r.

Omówienie

Artykuł przedstawia przykład analizy dany spektrometrycznych typu MALDI-TOF. Dane, po procesie usuwania wartości odstających i wstępnej analizie poddane zostały dekompozycji z wykorzystaniem widma średniego. Modelowanie oparte zostało na mieszaninach rozkładów normalnych, których parametry wyznaczono z użyciem algorytmu EM.

Tak przygotowane dane poddano klasyfikacji metodą SVM. Z uwagi na bardzo wysoką wymiarowość danych konieczne było użycie metod redukcji i ekstrakcji cech (PLS, SVM-RFE, test T). Wyniki klasyfikacji zostały poddane interpretacji biologicznej. Do tego celu posłużyła aplikacja autora, która zintegrowała dostęp do kilku biologicznych baz danych. Dzięki temu możliwa była kompleksowa analiza danych w badanym zbiorze.

Address

Małgorzata PLECHAWSKA-Wójcik: Politechnika Lubelska, Instytut Informatyki, ul. Nadbystrzycka 36B, 20-618 Lublin, Polska, gosiap@cs.pollub.pl.