STUDIA INFORMATICA 2010
Volume 31 Number 2A (89)

Małgorzata PLECHAWSKA
Politechnika Lubelska, Instytut Informatyki

# APPLICATION OF GAUSSIAN MIXTURE MODEL AND PROTEOMIC DATABASES IN THE MASS SPECTRA ANALYSIS – ARCHITECTURE OF SOFTWARE OF COMPREHENSIVE MASS SPECTROMETRY DATA PROCESSING

**Summary**. This paper presents assumptions of the protein mass spectra analysis software. All the spectra are modeled with Gaussian Mixture Models. Estimation of model parameters is done by means of an Expectation-Maximization algorithm. The obtained parameters are used for a further biological analysis. The software is integrated with four huge protein databases available on-line. The biological information about proteins may be achieved on the chosen level of some detail.

**Keywords**: GMM, mass spectrometry, proteomic knowledge base

# ZASTOSOWANIE MIESZANIN GAUSSOWSKICH ORAZ PROTEOMICZNYCH BAZ DANYCH W ANALIZIE WIDM MASOWYCH – ARCHITEKTURA SYSTEMU INFORMATYCZNEGO KOMPLEKSOWEGO PRZETWARZANIA DANYCH SPEKTROMETRYCZNYCH

**Streszczenie**. Artykuł przedstawia założenia systemu służącego analizie widm masowych białek. Widma są modelowane za pomocą mieszanin rozkładów normalnych. Oszacowanie ich parametrów oparte jest na algorytmie Expectation-Maximization. Uzyskane parametry są poddawane dalszej analizie biologicznej. System jest zintegrowany z czterema dużymi białkowymi bazami danych dostępnymi on-line. Informacja biologiczna dotycząca zawartych w widmie białek może być uzyskana na wybranym stopniu szczegółowości.

**Słowa kluczowe**: mieszaniny rozkładów normalnych, spektrometria masowa, białkowe bazy danych

## 1. Introduction

The presented project is dedicated to a mass spectrometry data analysis. The mass spectrometry is a widely used technique of tissue samples processing. It performs a samples ionization. Ions obtained in this process are accelerated by an electric field and finally hit in a detector. The goal of a mass spectrometry is to measure the mass due to charge (m/z) values. M/z values and ions intensities are presented in form of the mass spectrum. This technique may be applied in the determination of compounds isotopic composition, compounds identification and compounds mixtures, compounds structure determination. In some medical and biological application mass spectrometry can be used for early diagnosis, monitoring disease progression or treatments effects [1].

Mass spectrometry techniques differ in types of ion source and mass analyzers. The MALDI-TOF [4] technique is widely applicable in a proteomic research. The MALDI (*Matrix-Assisted Laser Desorption/Ionization*) is known as a soft ionization technique. These samples are mixed with a highly absorbing matrix, which, when bombarded with a laser, stimulates process of laser energy transformation into the excitation energy [2]. After this process analyte molecules are sputtered and spared. The TOF (*time of flight*) is a type of detector which determines the mass of ions. It is done on the basis of time of particular ions drift through the spectrometer. Obtained in a such way velocities and intensities [2] of ions are proportional to the mass-to-charge (m/z) ratio.

The row mass spectra obtained from the spectrometer need before analysis some preprocessing steps. Reliable results of analysis may be achieved after use of a set of steps involving: binning, interpolation, normalization, baseline correction, normalization, denoising, peaks detection [3] and alignment [2]. The sort of used preprocessing techniques depends on peaks detection and quantification method. One of the most important preprocessing steps is denoising, especially a baseline correction. Baseline is a special case of noise, intensifying especially in initial part of the spectrum, where M/Z values are low. Removal of this kind of noise flattens and averages the spectrum. The baseline correction is essential for further analysis and improves the quality of it. It is usually performed with multiple shifted windows with defined width. Normalization and interpolation are useful techniques helpful analyzing and comparing few spectra simultaneously. Interpolation is useful during the unification of measurements points [5] along with m/z axis of all spectra. Normalization [4] is scaling all spectra to a single value of area under the curve. This scaling is usually done forthe total ion current (TIC) value or for the constant noise.

The MALDI-TOF technique has an extensive application in the proteomic research. Spectrometry data usually has a form of mass spectrum, implementing dependency M/Z values (horizontal axis) and ions intensities (vertical axis). Proteomic data are of a great importance for the problem of an early-stage cancer detection. MALDI, as well as SELDI (*Surface Enhanced Laser Desorption/Ionization*) provides high-resolution measurements. However, a problem exists is high-dimensional data sets which need to be processed in a reasonable time.

In general, data sets should be analyzed as follows: preprocessing, peak detection, identification of biomarkers, classification [6]. There are some methodologies proposed for markers identification. In [7,8] were presented methodologies based on data binning, whereas in [9,10,6] were considered mass spectrometry peaks to represent biological information.

Mass spectra analysis has an extensive reflection in the literature. Peaks identification process may be performed with usage of comparison of apex and surrounding noise level [11], minimal value of true peak area threshold [9,10,2] or local maxima and choose peaks higher than a noise-proportional, non-uniform threshold[12,13]. Also a mean spectrum may be applied [14] as well as peak clusters [15]. Peaks was also decomposed with a sum of constituent functions [16]. Dijkstra [17], Kempka [18] and Noy & Fasulo [19] proposed also a decomposition with mixture models. Those papers however do not use mixture models in the context of a biomarkers search.

Another issue which needs to be stressed is a problem with a normalization and a redundancy of biological data. There are some proteomic and biological databases, which offer access to a more or less structured data. However, proteomics is quite a young filed of science and data gained by different research centers were devoid of a general enforced structure. Such structures are created and introduced nowadays, but it is a hard task which needs a lot of time to be accomplished. That is why different databases have the same data labeled with different standards. This makes usage of those data especially difficult in the context of a software development.

## 2. Methods

The method used for spectra analysis consists of Gaussian Mixture Model [20] (GMM)decomposition.. The single spectrum is modeled with the GMM where each peak representation is a particular Gaussian distribution. A proper determination of GMM components gives a precise spectrum decomposition. Obtained results (parameters of peaks) may be used for a biological context determination.

This mixture model is defined as a combination of a finite number of single probability distributions. The mixture model is defined with eq. (1).

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

where $K$ is the number of mixture components, $p_k, k = 1,2,...K$ describes mixture parameters including weights ($\alpha_k, k = 1,2,...K$, $\sum_{k=1}^{K} \alpha_k = 1$).

Set of parameters in the mixture depends on types of included distribution. In case of Gaussian distribution it is a mean $\mu_k$ and a standard deviation $\sigma_k$. Each mixture component is given also with weight, which represents its share in the whole mixture.

Number of mixture parameters depends on a number of mixture components. Solving complex mixtures are needed for an algorithm, which is able to estimate parameters in a reasonable time with a defined accuracy. An Expectation - Maximization (EM) algorithm is one of well known methods of a hidden and unknown parameters estimation. Gaussians are a proper type of a distribution, because they are suitable for a spectra noise modeling.

The EM algorithm is an iterative, nonlinear method. It is composed of two steps: an Expectation (E) and a Maximization (M) performed in a loop. A step E (eq. 2) calculates a conditional probability of belonging of sample $x_n$ to k^th component. A step M is the most time consuming part of EM and it consists in calculation of new parameters values. Eq. 3 presents a form of M step adjusted to the Gaussian distribution.

$$p(k \mid x_n, p^{old}) = \frac{\alpha_k^{old} f_k(x_n, p^{old})}{\sum_{\kappa=1}^{K} \alpha_k^{old} f_k(x_n, p^{old})} \tag{2}$$

$$\begin{aligned}
\mu_k^{new} &= \frac{\sum_{n=1}^{N} x_n p(k \mid x_n, p_{old})}{\sum_{n=1}^{N} p(k \mid x_n, p_{old})}, k = 1,2,...,K \\
(\sigma_k^{new})^2 &= \frac{\sum_{n=1}^{N} (x_n - \mu_k^{new})^2 p(k \mid x_n, p_{old})}{\sum_{n=1}^{N} p(k \mid x_n, p_{old})}, k = 1,2,...,K \\
\alpha_k^{new} &= \frac{\sum_{n=1}^{N} p(k \mid x_n, p^{old})}{N}
\end{aligned} \tag{3}$$

The EM algorithm in presented form uses only one dimensional data. Application EM algorithm to two dimensional mass spectrometry data needs proper modifications. A weighted version of EM [21, 22], was adjusted to the MS data. The second dimension represents intensities, which may be described as repeats of corresponding m/z values. Therefore, appropriate multiplications needs to be performed. A weighted version of the EM remains the E step unchanged (eq. 2), because the value of probability that a sample $x_n$ belongs to k^th component does not change. So, independently of number of the $x_n$ repetition in a vector, calculated probability is always the same. The calculations of the M step need to be performed in a slightly different way. Proper formulas are presented in eq. 4.

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

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

$$\alpha_k^{new} = \frac{\sum_{n=1}^{N} p(k \mid x_n, p^{old}) y_n}{N} \tag{4}$$

The EM algorithm is needed for initial values. It is essential to choose proper initial values, because it reduces time needed for calculations and probability of local maximum problem occurrence. Stopping in a local maximum may lead to an inaccuracy and a poor repeatability of results.

Two more issues concerns the number of mixture components and the stop criterion. The number of mixture components needs to be known before the run of the algorithm. It may be estimated with a Bayesian Information Criterion (BIC) [23]. This criterion is based on a value of a likelihood function. The optimal number of criterion should maximizes eq. (5)

$$-2 \cdot \ln p(x \mid k) \approx BIC = -2 \cdot \ln L + k \ln(n) \tag{5}$$

where $x$ is a sample of size $n$ and $k$ is the number of estimation parameters. $p(x \mid k)$ is the likelihood of the observed data. $L$ is the maximized value of the likelihood function for the estimated model.

To estimate the number of components differently we can use a standard peaks detection method. Such method based on fast finding local maxima and minima works. Their results may be used as a support in the initial parameters calculation and estimation of number of components. The choice of convergence criterion and accuracy is equally important. Criteria may be based on consecutive values of parameters or likelihood function and it may be calculated with a chosen distance. The use of the likelihood function (eq. 5) enables an employing maximum likelihood rule which states, that the higher value of likelihood function is, the better parameters estimation can be gained. The usage of the maximum likelihood rule gives a certainty of the stability, because of the monotonicity of the likelihood function. A broader view of the stop criterion problem is available in [24].

## 3. Software project

This article presents a project of the mass spectra analysis software. The idea of the system is to enable a decomposition of a single or a multiple spectra. There also should be possible to perform a biological analysis of achieved results. Therefore, presented project is divided into two parts. The first one is dedicated to the spectra analysis. The second one is responsible for a biological context of achieved information.

Both modules are indented to work independently. They might exchange data or be used as two separated programs (fig. 1).

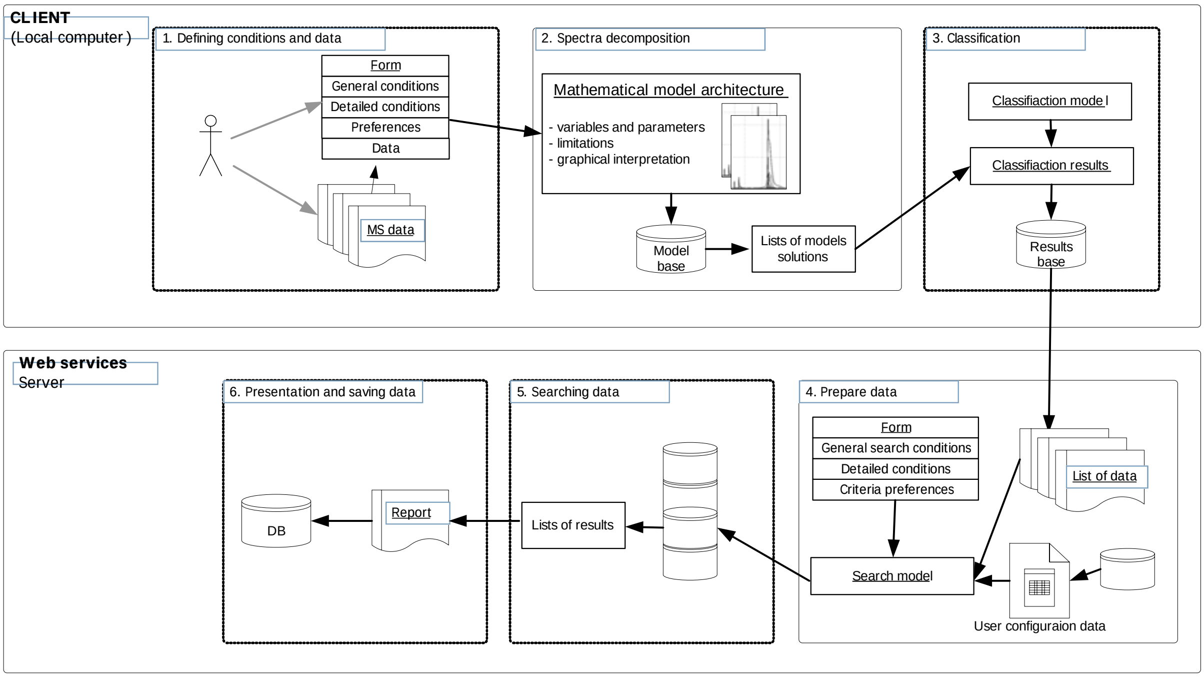

Fig. 1. A flowchart presenting data processing in the system
Rys. 1. Schemat przedstawiający przetwarzanie danych w systemie

### 3.1. Project of spectra analysis software

The spectra analysis module is designed as a standalone application. It is composed of a main decomposition module and a few associated sub-modules. The first level of a data flow diagram of the spectra analysis module is presented at fig. 2.

The main function of this module is the spectra decomposition with the GMM mixture model. This function performs one of two possible scenarios. The first one is processing all spectra sequentially, one by one using the method described in a Section 2 of this article. However, processing many spectra this way lengthens their calculations time, especially when the defined accuracy is high. In such cases the second method could be applied. This method performs calculations using the mean spectrum.

The mean spectrum calculation, compared to single spectrum decomposition, needs two more preprocessing steps. They are: interpolation, which standardizes m/z values (X axis) and normalization to Total Ion Current, which is needed for a spectra adjustment. A mean spectrum is calculated with eq. 6:

$$y\_sr_i = \frac{\sum_{j=1}^{K} y_{ij}}{K}, j = 1,2,...,N \tag{6}$$

where $y\_sr_i$ indicates i^th element in the output vector of mean values. $y\_sr_i$ corresponds to i^th element $x_i$. A X axis, after interpolation process, is uniformed for all analyzed spectra. $y_{ij}$ represents the i^th element of the intensities vector of j^th spectrum. $K$ is the number of spectra in the data-set and $N$ is the uniformed size of all spectra.

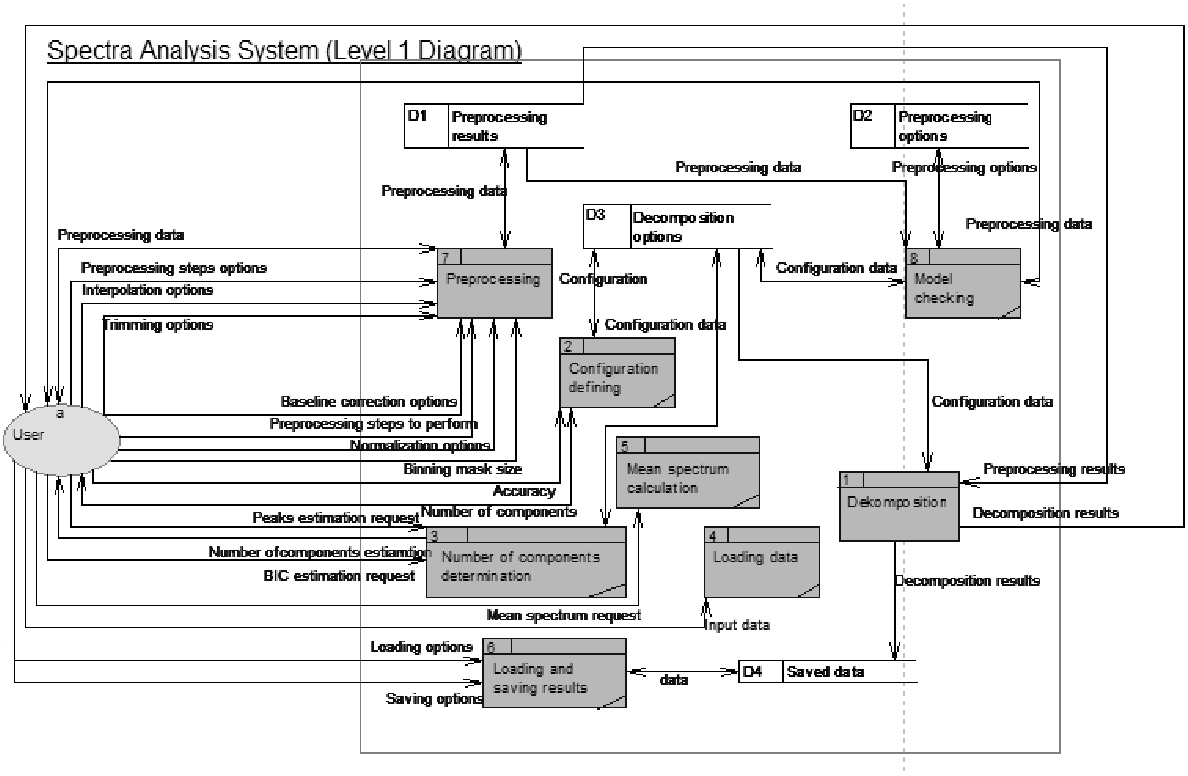

Fig. 2. DFD diagram of the spectra analysis module
Rys. 2. Diagram DFD modułu analizy widm

The mean spectrum is modeled with the GMM and decomposed with the EM algorithm. Obtained values of means ($\mu_k$) and standard deviations ($\sigma_k$) remain constant for every single spectrum in the set. Weights, however, needs to be obtained separately for each spectrum. The weight estimation is performed with widely known Least Squares methods.

Decomposition can be done after few steps of a data preparation. It is done by additional modules of the software. Data in form of a textual data are loaded and prepared in a Loading data module. The next step is a configuration. Settings involve a specification of :

- a special type of decomposition (with or without mean spectrum),
- a source and range of the EM initial data (randomization or with support of standard peaks determination algorithm),
- the accuracy of calculations,

- the number of components,
- preprocessing options (such as binning mask, baseline correction window, normalization and interpolation options, trimming and propriety checking).

Number of components may be defined directly but also a module of= a components determination number could be used. This module performs support of a standard peaks determination algorithm and also enables the BIC simulation. The last method last longer, but it gives clear and precise results.

The model checking module is designed to validate and control system settings before or at the beginning of decomposition. It enables viewing results of preprocessing performed by the Preprocessing module and checks possibility of use of defined settings. In particular, it enables checking baseline correction results and gives information about initial distribution of mixture model components.

The mean spectrum calculation module enables checking of the mean spectrum shape after defined preprocessing steps.

Results of the decomposition may be saved and transferred to the second module called a Biological context module.

### 3.2. Project of biological context module

This project concerns an application supporting results analysis. Results may be transferred directly from the spectra analysis software or they may be entered manually. The project assumes the possibility of passing through a four-steps path. Each step enables displaying data with a different level of biological context. Levels need to be achieved sequentially. The user may stop at any level and may go back to all previously generated levels. The diagram in Fig. 3 presents details of each level.

The Level0 is responsible for loading data and it gives the possibility of user data uploading. There is a possibility of choosing particular MS data transferred from the spectra analysis software. At this level the user may give some additional information specifying detailed search criteria. Those criteria includes: species, the MS platform, the possibility of double and triple charges. These criteria may be inserted manually or loaded from the user profile. Registered users are also provided with the possibility of saving results and history viewing as well as adding comments and sending data to the other users.

The Level1 is the m/z value level. The data chosen in Level1 are matched for further analysis. The m/z values are sent to an EPO-KB (Empirical Proteomic Ontology Knowledge Base) database [25]. The EPO-KB, after applying defined criteria, returns names of the found proteins. These results are obtained with a defined percentage tolerance. The results are presented for one spectrum at the time with references for further analysis and links to original results in the EPO-KB service. References and links concerns single proteins. The original results include particular detailed information about a specific protein.

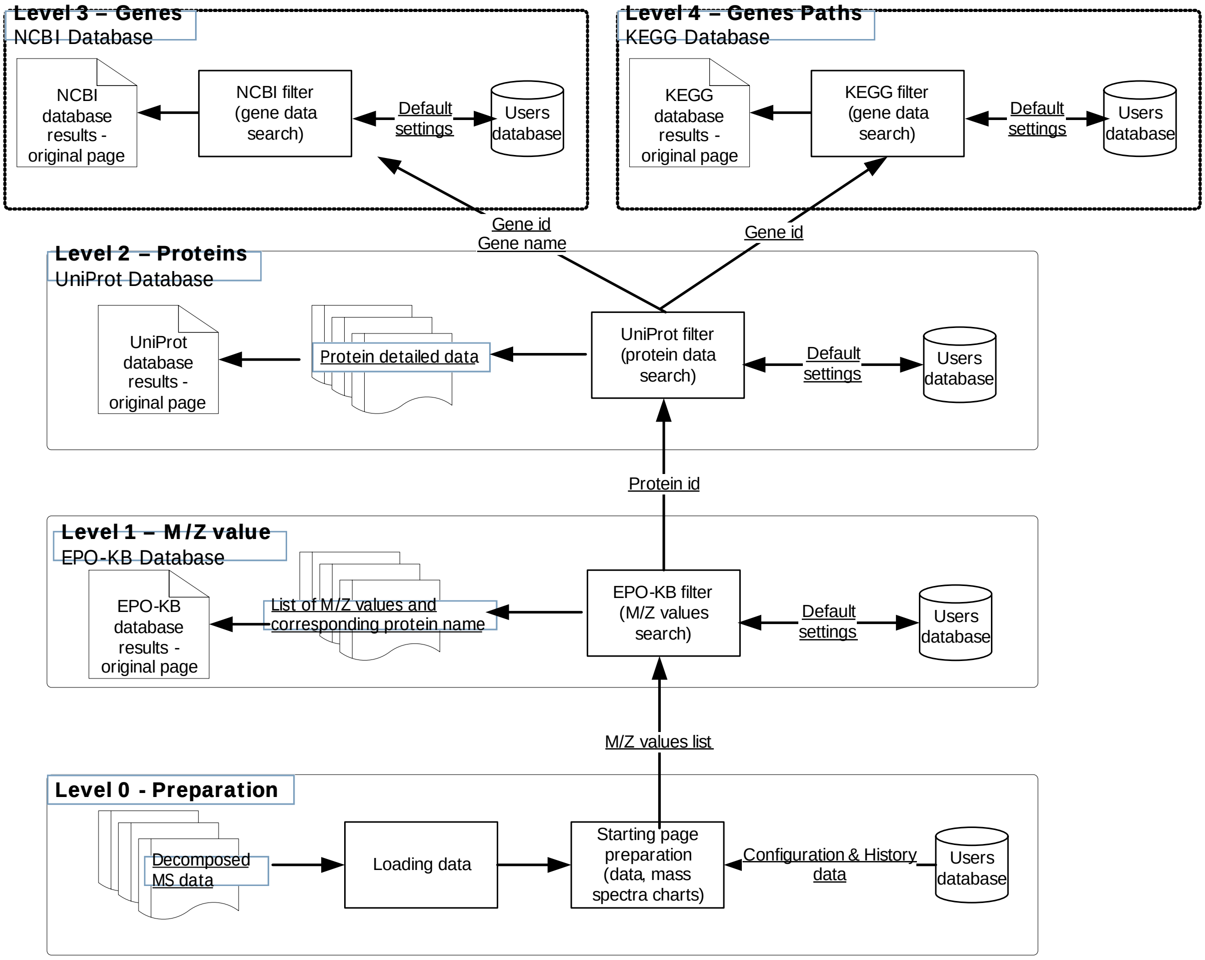
Level 3 – Genes
NCBI Database
NCBI database results - original page
NCBI filter (gene data search)
Default settings
Users database
Level 4 – Genes Paths
KEGG Database
KEGG database results - original page
KEGG filter (gene data search)
Default settings
Users database
Gene id
Gene name
Gene id
Level 2 – Proteins
UniProt Database
UniProt database results - original page
Protein detailed data
UniProt filter (protein data search)
Default settings
Users database
Protein id
Level 1 – M/Z value
EPO-KB Database
EPO-KB database results - original page
List of M/Z values and corresponding protein name
EPO-KB filter (M/Z values search)
Default settings
Users database
M/Z values list
Level 0 - Preparation
Decomposed MS data
Loading data
Starting page preparation (data, mass spectra charts)
Configuration & History data
Users database

Fig. 3. Flowchart of the biological context module
Rys. 3. Schemat działania modułu informacji biologicznej

The EPO-KB is a knowledge base containing biomarkers linked to proteins, peptides and their modifications [26]. This database contains also information about associated diseases taken from the literature. Proteins and peptides are associated with a range of m/z values. Additional conditions like post-translational modifications, biofluids, the MS method result in the necessity of some defining range of m/z values instead of a single one. The EPO-KB is based on an OWL ontology [epo1], which represents a knowledge linking m/z values to peptides and proteins on the MALDI and SELDI platforms. The knowledge base enables also entries of the additional *m/z* ratio to protein links. The ontology developed in the EPO-KB includes such elements as an ionized method, a type of a mass spectrometer, a type of a matrix and a biological sample, a laser power and charging of proteins (single, double or triple).

The Level2 of the biological context module is a protein level. This level is available after choosing proteins in the Level1. Detailed information about specific protein is obtained from an UniProt database. Data are filtered by such searching criteria as a name of the protein and, if needed, by a species type. Displayed results contains detailed information about proteins, such as entry name, status of reviewing process, organism, gene names and identifiers, features, GO annotations. More specific information are available on original UniProt page of the specific protein. Links to those pages and references for further analysis paths are also presented.

The UniProt database [27,28], produced by the UniProt Consortium, is repository of a protein sequence and an annotation data. Information are taken from the literature and other biological database and resources. The UniProt has implemented extended mechanisms of the data integration, standardization and redundancy avoidance. It is updated every 3 weeks. Information are available in XML and RDF formats. UniProd is composed from four main modules: the UniProt Archive, the UniProt Knowledgebase, the UniProt Reference Clusters and the UniProt Metagenomic and Environmental Sequence Database. Some of those databases have manually annotated records with proper literature information, the other have implemented some automated mechanisms. The database is extended and it covers such information as proteins, enzymes, biologically relevant domains and sites, post-translational modifications, subcellular locations, tissue specificity, splice isoforms, associated diseases or abnormalities as well as a history of all protein sequences and a metagenomic data.

The Level3 of the biological context module is a genes level. The Level2 enables displaying detailed information about genes coding a particular protein chosen at an earlier level. Those information are presented through a NCBI service. Searching may be based on a gene name or a gene identifier, whereas the gene identifier enables to return more accurate and precise information about a particular gene, its role, status, lineage and related data. From this level there is also possibility to get a gene pathways data from the KEGG database. It is the level4 of the module. User may obtained such information as details of a pathway, a structure, sequences, references to other databases.

The National Center for a Biotechnology Information (NCBI) [29] is an organization providing huge databases available online through the Entrez search engine. It provides the integrated database giving an access to a variety of genomes, sequence maps, chromosomes, integrated genetic and physical maps. This service enables finding organized information about organization of genes, its mappings as well as a comprehensive data about its structure, location, activities and association with a disease. It also makes the searching of biomarkers possible. It is composed of such databases as the GenBank DNA database with a widely known sequence comparison algorithm – BLAST. Other popular databases are: GEO (A Gene Expression and Hybridization Repository), dbSNP (Database of Nucleotide Sequence Variation) or Macromolecular Structure Databases. There are also available mechanisms of the data flow and processing as well as querying and linking the data.

The KEGG (Kyoto Encyclopedia of Genes and Genomes) [30] is a database of genomic, chemical and systemic functional information. It integrates plenty of tools and entry points such as BRITE ontologies, GENES databases or LIGAND database of chemical compounds and reactions. KEGG PATHWAY, responsible for mapping genomic or transcriptomic content of genes to the KEGG is provided also for global map of metabolic pathways. The KEGG, similar to other mentioned databases and services is available online and gives an access to resources through API, XML and FTP.

## 4. Summary

The presented project is a comprehensive bioinformatics tool enabling spectra pre-processing and analyzing. The use of the Gaussian Mixture Model decomposition enables a particular work with different types of spectra and options settings support minimizes risk of improper parameters selection. This method is slower than common spectra analysis techniques based on the local maxima and minima, although the computational time is acceptable. The second part of the system helps in extraction of biological information. It enables collecting essential information in one place with the possibility to save results.

**BIBLIOGAPHY**

1. Coombes K., Baggerly K., Morris J.: Pre-processing mass spectrometry data, Fundamentals of Data Mining in Genomics and Proteomics. W Dubitzky, M Granzow, and D Berrar, eds. Kluwer, Boston 2007, p. 79÷99.
2. 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 Vol. 21, No. 9, 2005, p. 1764÷1775.
3. 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, USA 2007.
4. Polanski A., Polanska J., Pietrowska M., Rzeszowska J., Stobiecki M., Tarnawski R., Skladowski K., Widlak P.: Application of the Gaussian mixture model to proteomic MALDI-ToF mass spectra. Journal of Computational Biology, Gliwice 2007.
5. 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.

6. Fushiki T., Fujisawa H., Eguchi S.: Identification of biomarkers from mass spectrometry data using a "common" peak approach. BMC Bioinformatics Vol. 7, 2006, p. 358.
7. Yu JS., Ongarello S., Fiedler R., Chen XW., Toffolo G., Cobelli C., Trajanoski Z.: Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. Bioinformatics Vol. 21, 2005, p. 2200÷2209.
8. Geurts P., Fillet M., de Seny D., Meuwis MA., Malaise M., Merville MP., Wehenkel L.: Proteomic mass spectra classification using decision tree based ensemble methods. Bioinformatics Vol. 21, 2005, p. 3138÷3145.
9. Yasui Y., Pepe M., Thompson ML., Adam BL., Wright GL. Jr., Qu Y., Potter JD., Winget M., Thornquist M., Feng Z.: A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. Biostatistics Vol. 4, 2003, p. 449÷463.
10. Tibshirani R., Hastie T., Narasimhan B., Soltys S., Shi G., Koong A., Le QT.: Sample classification from protein mass spectrometry, by 'peak probability contrasts'. Bioinformatics Vol. 20, 2004, p. 3034÷3044.
11. Eidhammer I., et al.: Computational methods for mass spectrometry proteomics. Wiley 2007.
12. Mantini D., et al.: LIMPIC: A computational method for the separation of protein signals from noise. BMC Bioinformatics Vol. 8, 2007, p. 101.
13. Mantini D., et al.: Independent component analysis for the extraction of reliable protein signal profiles from MALDI-TOF mass spectra. Bioinformatics Vol. 24, 2008, p. 63÷70.
14. McLachan G. J., Peel W.: Finite Mixture Distributions, Wiley 2000.
15. Zhang S.Q., et al.: Peak detection with chemical noise removal using Short-Time FFT for a kind of MALDI Data. Proceedings of OSB 2007, Lecture Notes in Operations Research Vol. 7, p. 222÷231.
16. Randolph T., et al.: Quantifying peptide signal in MALDI-TOF mass spectrometry data. Molecular & cellular proteomics: MCP Vol. 4(12), 2005, p. 1990÷9.
17. Dijkstra M., Roelofsen H., Vonk R. J., Jansen R. C.: Peak quantification in surface-enhanced laser desorption/ionization by using mixture models, Proteomics Vol. 6, 2006, p. 5106÷5116.
18. Kempka M., Sjodahl J., Bjork A., Roeraade J.: Improved method for peak picking in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry, Rapid Commun. Mass Spectrom. Vol. 18, 2004, p. 1208÷1212.
19. Noy K., Fasulo D.: Improved model-based, platform-independent feature extraction for mass spectrometry. Bioinformatics Vol. 23, No. 19, 2007, p. 2528÷2535.
20. Everitt B.S., Hand D.J.: Finite Mixture Distributions. Chapman and Hall, New York 1981.

21. Plechawska M., Polańska J., Polański A., Pietrowska M., Tarnawski R., Widlak P., Stobiecki M., Marczak Ł.: Analyze of Maldi-TOF proteomic spectra with usage of mixture of Gaussian distributions. Man-Machine Interactions, Advances in Intelligent and Soft Computing, Springer 2009.
22. 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.
23. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. Vol. 6, 1978, p. 461.
24. Plechawska M.: Comparing and similarity determining of Gaussian distributions mixtures. SMI 2008. Polish Journal of Environmental Studies. wyd. Hard Olsztyn. Vol.17, No. 3B, 2008.
25. Lustgarten J.L., et al.: EPO-KB: A searchable knowledge base of biomarker to protein links. Bioinformatics Vol. 24(11), 2008, p. 1418÷1419.
26. Lustgarten, J.L., et al.: Knowledge-based variable selection for learning rules from proteomic data. Bioinformatics Vol. 10(Suppl 9), 2009, p. 16.
27. UniProt Consortium: The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Vol. 38, 2010, p. D142÷D148.
28. 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 Vol. 10, 2009, p. 136.
29. Wheeler DL et al.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Vol. 37, 2009, p. D5÷D15.
30. 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 Vol. 36, 2008, p. D480÷D484.

Recenzenci: Dr inż. Dariusz Mrozek,
Prof. dr hab. inż. Andrzej Świerniak

Wpłynęło do Redakcji 31 stycznia 2010 r.

## Omówienie

Artykuł przedstawia założenia systemu służącego do analizy widm masowych typu MALDI-TOF. Główna metoda analizy widm polega na przeprowadzeniu dekompozycji w oparciu o mieszaniny rozkładów normalnych. Do oszacowania ich parametrów wykorzystywany jest algorytm Expectation-Maximization oraz metoda Największej Wiarogodności. Przed analizą właściwą dane muszą być poddane analizie wstępnej, która pozwoli na ich normalizację, uśrednienie, usunięcie z nich szumu.

Uzyskane w ten sposób wyniki mogą zostać poddane dalszej analizie. Polega ona na zbadaniu biologicznych właściwości próbki. Uzyskane w procesie analizy wartości m/z umożliwiają określanie składu białkowego próbki oraz uzyskaniu informacji o znalezionych białkach i genach je kodujących. System jest zintegrowany z czterema dużymi białkowymi bazami danych dostępnymi on-line. Informacja biologiczna dotycząca zawartych w widmie białek może być uzyskana na wybranym stopniu szczegółowości. Stopnie te obejmują: poziom wartości stosunku masy do ładunku, poziom białek, genów oraz ścieżek genowych.

## Address

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