MATHEMATICA APPLICANDA
Vol. 46(1) 2018, p. 77–86
doi: 10.14708/ma.v46i1.6384

Anna Glodek (Gliwice)
Joanna Polańska (Gliwice)

Method for deisotoping based on fuzzy inference systems.

Abstract Proteins are very significant molecules that can construct the fingerprint of cancer. When dealing with large molecules, such as proteins, the crucial issue is their trustful and precise identification. In the majority of cases, mass spectrometry is used to identify the protein. Processing of data gathered in mass spectrometry experiment consists of several steps, and one of them is deisotoping. It is an essential part of preprocessing because some peaks in the spectrum are not the unique compound, but they are members of an isotopic envelope. There are several existing methods of deisotoping, but none of them is general and can be used in any experimental settings. To manage this, we propose a new algorithm based on fuzzy inference systems. The method was tested on the data provided by Institute of Oncology in Gliwice, that has been gathered in MALDI experiment in two different settings on head and neck cancer tissue samples. The comparison study, done between the developed fuzzy-based algorithm and mMass method revealed that the proposed method was able to identify more consistent with the expert annotation isotopic envelopes.

2010 Mathematics Subject Classification: Primary: 92D40; Secondary: 92D30..

Key words and phrases: fuzzy logic, fuzzy inference systems, deisotoping, mass spectrometry, algorithms, MALDI ToF.

1. Introduction Nowadays, oncology is focused on identifying the proteins that could play a significant role in cancer diagnosis [13] and treatment efficiency evaluation. That is why it is highly important to properly identify, with the help of mass spectrometry, the proteins located in the cancer region. The mass spectrometer measures the ion masses, which form the peaks of the mass spectrum. One of the widely used mass spectrometry technique is matrix-assisted laser desorption and ionisation (MALDI) [5]. This method allows the detection of proteins, peptides, lipids and also exogenous and endogenous small molecules [7, 17] in tissue samples. It can be also used in mass spectrometry imaging experiments, which combine molecular evaluation of several analytes, the high sensitivity and selectivity of mass spectrometry with morphological information about the spatial distribution of molecules in tissues [7, 15]. In MALDI experiment the sample is loaded into the mass spectrometer, where it is ionised. Then it is pulled into the mass analyser where the molecules are separated based on their mass to charge ratio. The

detector records the charge induced when an ion passes a surface [1, 22]. The output data is a raw mass spectrum - mass over charge ratio and intensity of the peaks. Many articles list numerous advantages of using MALDI technique for cancer research [7, 15, 17]. To properly handle with such a data, several preprocessing methods have to be applied to obtain a data where unique spectral fragments are represented by only one datum named peaks [4]. One step of preprocessing is deisotoping, since some peaks in the spectrum are not the unique compound but there are members of an isotopic envelope, and it turns out that there are the isotopes of one compound. To perform deisotoping, the members of an isotopic envelope should be identified. Then, the isotopic envelope could be reduced to one peak that is described by the mass of the first monoisotopic peak in the isotopic envelope and the intensity is equal to the sum of intensities of all member peaks. Deisotoping is a very crucial step in spectrum preprocessing, leading to the significant reduction of redundant data. Thus, the protein identification is more accurate.

There is plethora of existing deisotoping methods, but each of them is dedicated to either high-resolution mass spectra or low-resolution. The high-resolution mass spectrometer can measure ion masses very accurately, and it can detect the minute differences in mass between two compounds, whereas, on a low-resolution mass spectrometer, the masses would appear to be identical [1]. Usually, the methods are dedicated to the specific molecules and the particular kind of mass spectrometry experiment, for instance, MALDI (Matrix-assisted laser desorption and ionization) [5], LC-MS (Liquid chromatography-mass spectrometry) [12], ESI (Electrospray ionization) [6], EI (Electron Ionisation) [2] etc. For example, YADA is dedicated to high-resolution mass spectra for large peptide molecules. It takes into consideration overlapping isotopic envelopes. It filters noise peaks and then discards peaks that do not contribute to charge determination - the intensities of peaks will monotonically increase until a local maximum is achieved. It has been tested only on LC-MS data [3]. mMass can be used for different kinds of spectra, but to perform deisotoping, one must annotate the peaks firstly. Algorithm takes into consideration mass shift between the isotopes and theoretical intensity that is calculated using the averagine formula. [20]. Once all the peaks are labelled, the algorithm removes unwanted isotopes from the final peak list. MS-Deconv can deisotope complex mass spectra and takes into account overlapping isotopic envelopes but it is only introduced for proteins, and the tests were based on top-down proteomics - LC-MS data [8]. This combinatorial algorithm firstly generates a large set of candidate envelopes, constructs an envelope graph encoding all envelopes and relationships between them and then finds the heaviest path in the envelope graph. [8]. BPDA can be used only for MALDI-ToF and LC-MS experiments for high-resolution mass spectra for only proteins, and it is based on Bayesian approach. It looks for all possible combinations of possible peptide candidates and iteratively finds

the best fitting peptide parameters to minimise the mean squared error of the inferred spectrum to the observed one [21]. LipidQA is dedicated only to lipids and LTQ and Q-TOF experiments. It is based on calculating the theoretical isotope distributions by comparing MS/MS spectra obtained in a data-dependent manner to a library of reference spectra of complex lipids [19]. Performing deisotoping process is difficult because there are many possible ways of resolving the problem and, although many existing methods, there is none general one applicable to all mass spectrometry techniques. There also no benchamrk datasets, that would allow comparison across different mass spectrometry settings. We propose a fuzzy-based approach featuring high generality, flexibility and possibility of introduction the expert knowledge to if-then rules.

2. Methods The fuzzy system, that works on Mamdani-Assilan model has been developed. It is based on if-then fuzzy rules according to the formula (1) [18]

$$R = \{R^{(i)}\}_{i=1}^{I} = \{ \text{ if } (\bigwedge_{n=1}^{N} X_n \text{ is } A_n^{(i)}), \text{ then Y is } B^{(i)}\}_{i=1}^{I} \tag{1}$$

where: $X_1, X_2, ..., X_N$ - input linguistic variables; Y - output linguistic variable; $A_1^{(i)}, A_2^{(i)}, ..., A_N^{(i)}, B^{(i)}$ - linguistic values for i-th rule.

The rules based on linguistic variables for our fuzzy-based system for deisotoping are as follows:

1. If the distance between two neighbouring peaks is approximately equal to 1 Dalton [Da], then a peak is a member of an isotopic envelope.
2. If variance ratio of two neighbouring peaks is approximately equal to one, then a peak is a member of an isotopic envelope.
3. If an amplitude ratio between two neighbouring peaks is decreasing, then the peaks are the members of an isotopic envelope.

Mamdani and Assilan used the minimum operation as t-norm that models AND connectors in if-then rules and also as conjunction interpretation of these rules. For the aggregation of the rules the maximum operator is used, while for defuzzification - the centre of gravity method [18].

The parameters for the membership functions have been set by thorough analysis of real isotopic envelope characteristic features that take into consideration abundance of isotopes and were calculated by widely used isotope pattern calculator. They have also been confirmed by the expert.

An element is called [18]:

1. Not included in the fuzzy set (membership function equals zero - equivalent of not being a member of a crisp set)

2. Fully included (membership function equals 1 - the equivalent of being a member of the crisp set)
3. Partially included (membership function $0 < \mu_A < 1$)

3. Results Data has been provided by Institute of Oncology in Gliwice, and it has been gathered in MALDI experiment based on head and neck cancer data. Into this research were enrolled 120 male cancer patients: 35 patients with squamous cell cancer located in head and neck region (samples were analysed using an Autoflex MALDI-ToF mass spectrometer (Bruker Daltonics, Bremen, Germany); the analyser worked in the linear mode, and positive ions were recorded in the mass range between 2 and 13 kDa [14]. The raw spectrum was firstly preprocessed (it includes baseline removal, noise filtering etc.). The peaks were identified using the Gaussian Mixture Modeling [10, 11]. After spectrum preprocessing, we performed deisotoping on the randomly chosen one peptide and one lipid dataset to present the properties of the developed algorithm. Peptides are molecules that consist of between 2 to 50 amino acids, and they are less defined in a structure in comparison with proteins [16]. Lipids make up the building blocks of the structure and function of living cells. These molecules contain hydrocarbons [9]. mMass algorithm was chosen for comparison with our fuzzy-based algorithm due to its ability to detect the isotopic patterns for both peptides and lipids and due to the fact that it also takes into account the mass shift between isotopes - that condition has been mentioned as a very important one by the expert in the field of mass spectrometry.

Peptides There were 492 isotopic envelopes found in total with 1249 peaks identified as members of the isotopic envelopes. The longest isotopic envelope consisted of 6 peaks (Table 1). Originally, there were 2328 peaks in a spectrum while after deisotoping left 1249 peaks in the spectrum that resulted in peak reduction by 33%.

While mMass algorithm was applied to the same spectrum, only 137 peaks have been classified as the members of 64 isotopic envelopes. There were 2 isotopic envelopes that consist of 5 peaks, 4 consist of 3 peaks etc., twelve peaks have been identified as the members of the 1-element isotopic envelope, and probably there are the false discoveries.

More than 58% of isotopic peaks identified by mMass were also identified by our approach (Figure 1). Comparison to the expert knowledge revealed that the common part of isotopic peaks that were classified by our fuzzy-based algorithm and by an expert was approximately 63%. The number decreases significantly in case of the mMass algorithm, where common number of peaks identified by expert and by mMass is approximately 35%.

Lipids Within the lipid spectrum, 493 isotopic envelopes have been detected constructed by 1309 peaks. The longest isotopic envelope consists of

8 peaks. The fuzzy-based deisotoping brought dimension reduction by 54%, keeping 1303 peaks of original 2398. The number of the isotopic envelopes detected by a fuzzy based algorithm is more than triple higher of mMass ones (Figure 2). Figure 2 presents the number of the isotopic envelope identified by fuzzy-based algorithm and mMass.

In comparison to mMass, the fuzzy-based algorithm can identify much more members of the isotopic envelopes. It also can detect overlapping isotopic envelopes. What is more, according to the expert knowledge in the field of mass spectrometry, isotopic envelopes identified by mMass are very often too long, since for example, the average lipid isotopic envelope consists from 2 - 8 peaks, whereas mMass has found an isotopic envelope with the length of 86 peaks. The fuzzy-based algorithm works quite accurately for both peptides and lipids, so there is no constraint of the specific type of molecules. There were examples of isotopic envelopes that have been identified by only one algorithm - either by mMass or fuzzy-based algorithm. The probable cause of such a situation is that the basis of work of each algorithm is quite different. The fuzzy-based algorithm takes into consideration intensity, the distance between neighbouring peaks and the ratio of variances of the neighbouring peaks, while mMass takes into account theoretical isotope pattern, and the intensity of each peak is compared with its theoretical intensity (Figure 3, Figure 4, Figure 5).

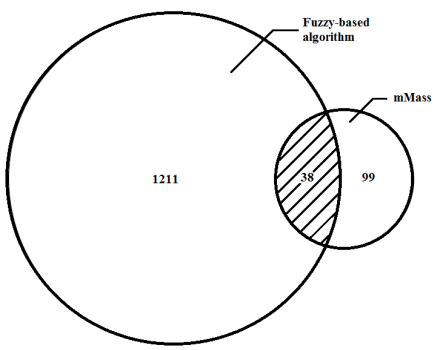

Figure 1: Comparison of the number of peaks deisotoped by fuzzy based algorithm vs mMass for peptides

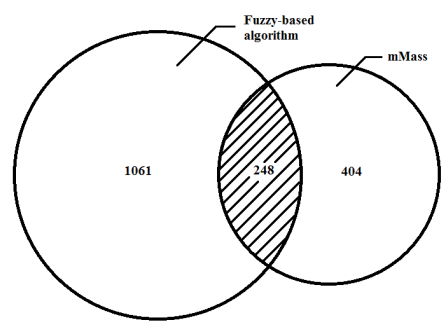

Figure 2: Comparison of the number of peaks deisotoped by fuzzy based algorithm vs mMass for lipids

4. Conclusion Our results show that fuzzy-based algorithm is more flexible in comparison with mMass. More isotopic envelopes consistent with the

	Peptides		Lipids	
Isotopic envelope length	**Fuzzy-based algorithm**	**mMass**	**Fuzzy-based algorithm**	**mMass**
	No. of isotopic envelopes (No. of peaks)			
$\geq$ 8	0 (0)	0 (0)	1 (8)	20 (408)
7	0 (0)	0 (0)	3 (21)	4 (28)
6	5 (30)	0 (0)	9 (54)	5 (30)
5	10 (50)	2 (10)	11 (55)	3 (15)
4	41 (164)	3 (12)	44 (176)	3 (12)
3	133 (399)	9 (27)	145 (435)	17 (51)
2	303 (606)	38 (76)	280 (560)	30 (60)
1	0 (0)	12 (12)	0 (0)	48 (48)
Total	**492 (1249)**	**64 (137)**	**493 (1309)**	**130 (652)**

Table 1: Number and length of isotopic envelopes obtained by fuzzy-based algorithm and mMass for peptides and lipids

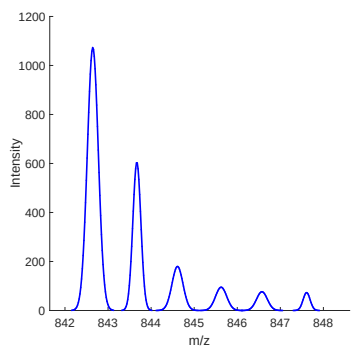

Figure 3: The exemplary isotopic envelope identified by the fuzzy-based algorithm

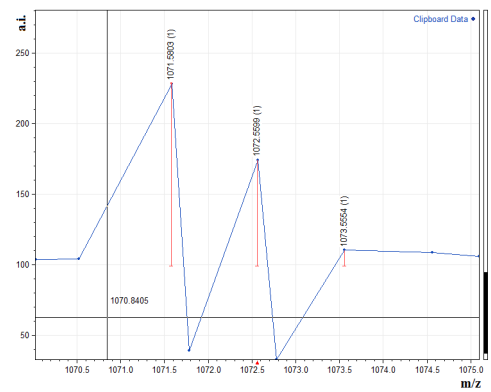

Figure 4: The exemplary isotopic envelope identified by mMass

expert knowledge have been identified by the fuzzy-based algorithm. It also successfully handles the overlapping isotopic patterns. As deisotoping is the vital part of mass spectra preprocessing, it results in the reduction of redundancy in data, allowing to identify the proteins more precisely and accurately. Our fuzzy-based algorithm can be widely used in different mass spectrometry settings, although it was presented here in the context of MALDI-ToF experiments only.

5. Acknowledgements The work was financially supported by Polish National Science Centre grant BiTIMS 2015/19/B/ST6/01736. The authors

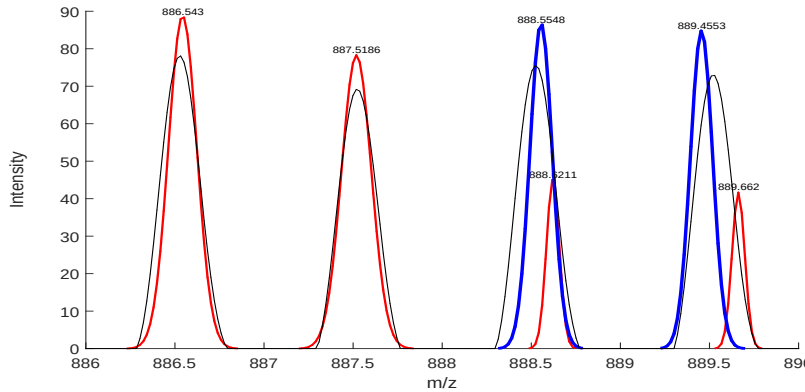

Figure 5: The exemplary overlapping isotopic envelopes identified by fuzzy-based algorithm (in red - first isotopic envelope, in blue - second isotopic envelope)

are very thankful to Jacek Łęski for his help in the design of fuzzy inference system and to Monika Pietrowska and Marta Gawin for their help in defining the isotopic patterns. We are also thankful to Katarzyna Bednarczyk for performing data preprocessing and GMM modelling.

References

[1] E. de Hoffmann and V. Stroobant. *Mass Spectrometry. Principles and applications*. Wiley, 2007. ISBN 978-0-470-03310-4. Cited on p. 78.

[2] W. Bleakney. A new method of positive ray analysis and its application to the measurement of ionization potentials in mercury vapor. *Phys. Rev.*, 34:157–160, 1929. doi: 10.1103/PhysRev.34.157. Cited on p. 78.

[3] P. C. Carvalho, T. Xu, X. Han, D. Cociorva, V. C. Barbosa, and J. R.Yates, III. Yada: A tool for taking the most out of high-resolution spectra. *Bioinformatics*, 25:2734–2736, 2009. doi: 10.1093/bioinformatics/btp489. PMID: 19684088 [PubMed] Cited on p. 78.

[4] I. Eidhammer, K. Flikka, L. Martens, and S. O.Mikalsen. *Computational Methods for Mass Spectrometry Proteomics*. Wiley, 2007. ISBN 978-0-470-51297-5. Cited on p. 78.

[5] F. Hillerkamp, M. Karas, R. C.Beavis, and B. T.Chait. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem*, 63:1193A–1203A, 1991. doi: 10.1021/ac00024a716. PMID: 1789447 [PubMed]. Cited on pp. 77 and 78.

[6] C. Ho, C. Lam, M. Chan, R. Cheung, L. Law, L. Lit, K. Ng, M. Suen, and H. Tai. Electrospray ionisation mass spectrometry: Principles and clinical applications. *Clin Biochem Rev.*, 24:3–12, 2003. PMID: 18568044 [PubMed]. Cited on p. 78.

[7] J. Kriegsmann, M. Kriegsmann, M., and R. Casadonte. MALDI-TOF imaging mass spectrometry in clinical pathology: A valuable tool for cancer diagnostics (review). *Int. J. Oncol.*. 46(3):893–906, 2015. doi: 10.3892/ijo.2014.2788. PMID: 25482502 [PubMed]. Cited on pp. 77 and 78.

[8] X. Liu, Y. Inbar, P. C. Dorrestein, C. Wynne, N. Edwards, P. Souda, J. P. Whitelegge, V. Bafna, and P. A. Pevzner. Deconvolution and database search of complex tandem mass spectra of intact proteins: A combinatorial approach. *Molecular & Cellular Proteomics*, 9(12):2772–2782, 2010. doi: 10.1074/mcp.M110.002766. PMID: 20855543 [PubMed]. Cited on p. 78.

[9] J. Berg, J. Tymoczko, and L. Stryer. *Biochemistry*, 7th edition. Palgrave MacMillan, 2011. ISBN 14-29-27635-5. Cited on p. 80.

[10] A. Polański, M. Marczyk, M. Pietrowska, P. Widłak, J. Polańska. Signal Partitioning Algorithm for Highly Efficient Gaussian Mixture Modeling in Mass Spectrometry. *PLOS ONE*, 10(7):e0134256, 2015. doi: 10.1371/journal.pone.0134256. PMID: 26230717 [PubMed] Cited on p. 80.

[11] A. Polański, M. Marczyk, M. Pietrowska, P. Widłak, and J. Polańska. Initializing the EM Algorithm for Univariate Gaussian, Multi-Component, Heteroscedastic Mixture Models by Dynamic Programming Partitions. *International Journal of Computational Methods*, 15(3):1850012 (21 pages), 2018. doi: 10.1142/S0219876218500123. Cited on p. 80.

[12] W. M. Niessen. *Liquid Chromatography-Mass Spectrometry*. Boca Raton: CRC Taylor & Francis., 3rd edition, 2006. ISBN 978-14-200-1454-9. Cited on p. 78.

[13] M. Pietrowska, H.C. Diehl, G. Mrukwa, M. Kalinowska-Herok, M. Gawin, M. Chekan, J. Elm, G. Drążek, A. Krawczyk, D. Lange, H.E. Meyer, J. Polańska, C. Henkel, and P. Widłak. Molecular profiles of thyroid cancer subtypes: Classification based on features of tissue revealed by mass spectrometry imaging. *Biochim. Biophys. Acta. - Proteins and Proteomics*, 1865(7):837–845, 2017. doi: 10.1016/j.bbapap.2016.10.006. PMID: 27760391 [PubMed]. Cited on p. 77.

[14] M. Pietrowska, J. Polańska, R. Suwiński, M. Wideł, T. Rutkowski, M. Marczyk, I. Domińczyk, L. Ponge, L. Marczak, A. Polański, and P. Widłak. Comparison of peptide cancer signatures identified by mass spectrometry in serum of patients with head and neck, lung and colorectal cancers: Association with tumor progression. *Int J Oncol*, 40:148–156, 2012. doi: 10.3892/ijo.2011.1186. PMID: 21894432 [PubMed]. Cited on p. 80.

[15] M. Rodrigo, O. Zitka, S. Krizkova, A. Moulick, V. Adam, and R. Kizek. MALDI-TOF MS as evolving cancer diagnostic tool: A review. *J Pharm*

Biomed Anal., 95:245-55, 2014. doi: 10.1016/j.jpba.2014.03.007. PMID: 24699369 [PubMed]. Cited on pp. 77 and 78.

[16] L. Eaton and K. Rogers. *Examining Basic Chemical Molecules (The Building Blocks of Life)*. Rosen Education Service, 2018. ISBN 978-15-383-0005-3. Cited on p. 80.

[17] C. Schöne, H. Höfler, and A. Walch. MALDI imaging mass spectrometry in cancer research: Combining proteomic profiling and histological evaluation. *Clin Biochem*, 46(6):539-45, 2013. doi: 10.1016/j.clinbiochem.2013.01.018. PMID: 23388677 [PubMed]. Cited on pp. 77 and 78.

[18] E. Czogała and J. Łęski. *Fuzzy and Neuro-Fuzzy Intelligent Systems*. Physica-Verlag, Heidelberg, 2000. ISBN 978-37-908-1289-3. Cited on p. 79.

[19] H. Song, F. Hsu, J. Ladenson, and J. Turk. Algorithm for processing raw mass spectrometric data to identify and quantitate complex lipid molecular species in mixtures by data-dependent scanning and fragment ion database searching. *J Am Soc Mass Spectrom.*, 18(10):1848-58, 2007. doi: 10.1016/j.jasms.2007.07.023. PMID: 17720531 [PubMed]. Cited on p. 79.

[20] M. Strohalm, M. Hassman, B. Kosata, and M. Kodicek. MMASS data miner: an open source alternative for mass spectrometric data analysis. *Rapid Commun Mass Spec*, 22(6):905-8, 2008. doi: 10.1002/rcm.3444. PMID: 18293430 [PubMed]. Cited on p. 78.

[21] Y. Sun, J. Zhang, U. Braga-Neto, and E. Dougherty. BPDA-a Bayesian peptide detection algorithm for mass spectrometry. *BMC Bioinformatics*, 11(490), 2010. doi: 10.1186/1471-2105-11-490. PMID: 20920238 [PubMed]. Cited on p. 79.

[22] R. S. Tuma. MALDI-TOF mass spectrometry: Getting a feel for how it works. *Oncology Times*, 25(19):26, 2003. doi: 10.1097/01.COT.0000290986.00178.61. Cited on p. 78.

Metoda identyfikacji obwiedni izotopowych oparta na systemach rozmytych.

Anna Glodek i Joanna Polańska

Streszczenie Praca przedstawia nowy algorytm identyfikacji obwiedni izotopowych w widmach proteomicznych MALDI ToF. W ostatnich latach proteomika, wraz z genetyką i transkryptomiką, silnie wspierają diagnostykę chorób nowotworowych. Bardzo ważne jest precyzyjne zidentyfikowanie białek znajdujących się w obszarze raka, gdyż pozwala to zrozumieć proces nowotworzenia oraz zaplanować właściwą terapię. Spektrometia mas, a właściwie technika zwana MALDI ToF (ang. Matrix Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry) jest powszechnie stosowana do pozyskania widm masowych, w których zawarta jest informacja o liczbie jonów o danym stosunku masy do ładunku. Etap przetwarzania wstępnego sygnału wymaga m.in. usunięcia szumu, linii bazowej i normalizacji. Identyfikacja

obwiedni izotopowych jest również niezwykle ważnym etapem procesu przetwarzania wstępnego, który pozwala na usunięcie redundancii i zredukowanie wymiarowości danych. Istnieje wiele algorytmów identyfikacji obwiedni izotopowej, jednak każdy z nich przeznaczony jest dla innego rodzaju techniki spektrometrii masowej (MALDI, LC-MS, ESI, etc.) bądź dla konkretnego rodzaju cząsteczek. Zaproponowany algorytm oparty jest na teorii systemów rozmytych, a reguły wnioskowania zostały opracowane we współpracy z zespołem ekspertów w dziedzinie spektrometrii masowej. Przetestowany został na danych uzyskanych z Instytutu Onkologii im. Marii Skłodowskiej-Curie w Gliwicach, pochodzących z badań nad rakiem głowy i szyi. Wyniki autorskiego algorytmu do identyfikacji obwiedni izotopowych porównano z jedną z istniejących metod do identyfikacji obwiedni izotopowych.

Klasyfikacja tematyczna AMS (2010): 92D40; 92D30.

Słowa kluczowe: matematyka stosowana, systemy rozmyte, logika rozmyta, proteomika, spektrometria mas, obwiednie izotopowe.

Anna Glodek Master degree in Automatic Control and Robotics, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology in 2015. PhD student in the field of Biocybernetics and Biomedical Engineering since 2015 in Silesian University of Technology.

Joanna Polańska PhD in the field of Automatic Control and Robotics in 1996, the Silesian University of Technology. Habilitation in 2008 in Biocybernetics and Biomedical Engineering, Nalecz Institute, Polish Academy of Sciences. Awarded the title of professor in 2015. Member of the Committee of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences. Data scientist, expert in bioinformatics, machine learning, and biostatistics. Listed in the European Mathematical Society, FIZ Karlsruhe, and the Heidelberg Academy of Sciences bibliography database known as zbMath under ai:Polanska.Joanna, in MathSciNet under ID:726680. She has records at ORCID ID:0000-0001-8004-9864 and Web of Science ResearcherID:H-4513-2013.

Anna Glodek
Silesian University of Technology
Faculty of Automatic Control, Electronics and Computer Science
Gliwice, Akademicka 16 44-100, Poland
E-mail: anna.glodek@polsl.pl

Joanna Polańska
Silesian University of Technology
Faculty of Automatic Control, Electronics and Computer Science
Gliwice, Akademicka 16 44-100, Poland
E-mail: joanna.polanska@polsl.pl

Communicated by: Urszula Foryś

(Received: 15th of May 2018; revised: 2nd of July 2018)