10.2478/v10170-010-0026-3

Archives of Control Sciences
Volume 20(LVI), 2010
No. 4, *pages 457–472*

Mixture model of NMR – its application to diagnosis and treatment of brain cancer

FRANCISZEK BINCZYK, RAFAŁ TARNAWSKI and JOANNA POLAŃSKA

Nuclear Magnetic Resonance (NMR) is widely used technique in cancer diagnosis and treatment planning. It is employed to search for the high concentration regions of particular metabolites, which are directly related to the concentration of cancer cells. NMR signal maybe be characterized by a set of peaks which are representation of every distinct metabolite. Area under peak must be calculated in order to obtain proper information about metabolite amount. Commercially available software allows for the analysis of one-peak-in-time only. The proposed technique, based on Gaussian Mixture Model (GMM), allows for modeling all-peaks-in-time, and corrects after the neighboring peaks giving more accurate estimates of metabolite concentration. The resulting software processes NMR signal from the very beginning up to the final result, which is given in a form of so called metabolite map.

Key words: GMM, EM algorithm, BIC, NMR, Savitzky-Golay filter

1. Introduction

Nuclear Magnetic Resonance (abbr. NMR) is a technique of detecting molecule amounts in tested specimens. This technique uses phenomenon of nucleus magnetic resonance. The nucleus is placed in a homogenous external magnetic field and a radio frequency signal is applied , causing variations in spin orientation of the nucleus. After signal application, the nucleus gets back to its original orientation radiating out energy in a process called precession. The emitted signal has its own frequency depending on neighborhoods electron clouds or connections between nuclei. For different metabolite concentrations, the obtained number of spectrum peaks also differs [2]. Thanks to the fact, that the signal is received with different precession frequencies, it can be visible in the Fourier domain as a set of peaks. Every distinct peak refers to one of the metabolites

F. Binczyk and J. Polańska are with Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland. R. Tarnawski is with Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Poland. E-mails: franciszek.e.binczyk@polsl.pl, tarnawski@io.gliwice.pl, joanna.polanska@polsl.pl.

We are very grateful to employees of Cancer Center and Institute of Oncology in Gliwice: Prof. Barbara Bobek-Bilewicz, Anna Hebda and Jakub Połetek, for providing help and access to the medical data. The work was partially supported by MNiSW, N N402 350638, 17.03.2010-16.03.2013the authors are very thankful to the reviewers for their suggestions and remarks that allow for manuscript improvement.

Received 18.11.2010. Revised 8.12.2010.

present in the tested specimen. An exemplary NMR signal is shown in Fig. 1. One of NMR's application is to diagnose tumors of the human brain [2]. In such application, a so-called tested area is defined. This space is divided into smaller parts called voxels. The voxel can be defined of a different size, typically it is $1 \times 1 \times 1$ cm. An exemplary voxel is presented in Fig. 2.

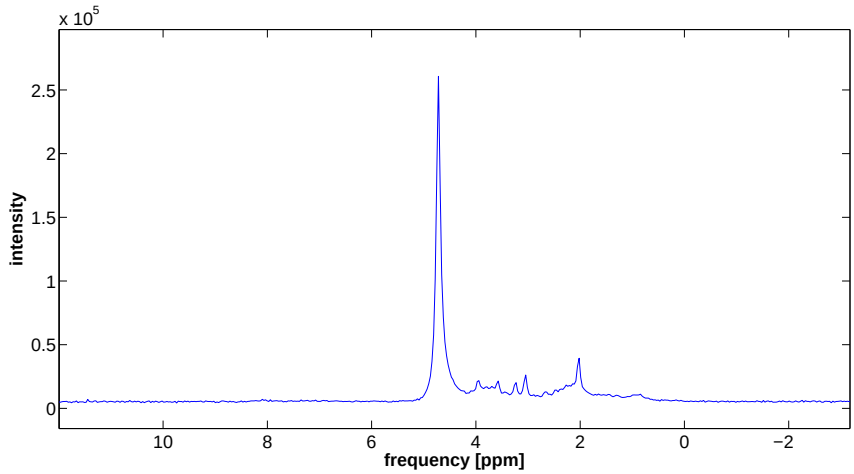

Figure 1. Exemplary signal obtained for 1H human brain spectroscopy.

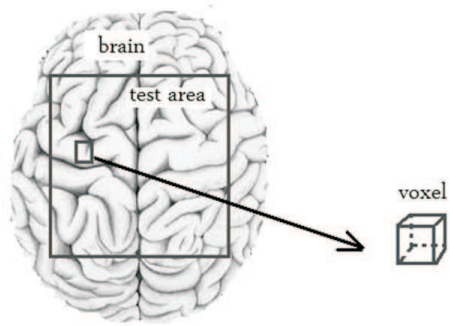

Figure 2. Exemplary voxel from test area of human brain.

There are several commercially available software allowing for the analysis of the NMR signal e.g. Tarquin [9], AQSES [12] or jMRUI [13]. All of them assume that only one, particular metabolite is under investigation. The concentration analysis of more than one metabolite calls neeeds repetition of the peak modeling. The list of available metabolites is limited as well. Thus, a new idea based on modeling a set of peaks as

a Gaussian mixture is proposed. This methodology is similar to the approach used by Pietrowska et al. [8] in case of MALDI ToF proteomic data analysis.

The multistep methodology is proposed and implemented in order to obtain quantitative analysis of NMR spectroscopic data. Developed software is able to analyze signals obtained from different spectrometers. It starts with raw data reading, then signal is preprocessed, which consists of the following, very important steps:

- transformation to frequency domain,
- signal normalization,
- removal of the dominating water/reference peak.

After the preprocessing is ready, signal modeling is performed. Final result, according to the demand, is presented in a form of the so-called 'metabolite map'. It means that particular metabolite amounts are calculated for every distinct voxel of the tested area and then the data is reordered according to its original position in the tested area of the human brain.

2. Signal preprocessing

2.1. Fourier transformation

Free induction decay (FID) is the signal that corresponds to the energy radiated from non-equilibrium nuclear spin magnetization in a constant and homogenous magnetic field (Fig. 3). Such a magnetization is obtained by applying an external pulse of given radio frequency. Magnetization caused by precessing nucleons induces oscillating voltage in a detector. Not every magnetization is able to induce voltage. An assumption that magnetization poses at least one non-zero component in x-y plane has to be satisfied.

Voltage detected by a surrounding coil is a time domain signal. To obtain resonance frequencies of magnetized nuclear spins it is necessary to apply the Fourier transform [1][11] to FID data.

2.2. Signal normalization

To acquire precise information about metabolite amounts in the tested area, consisting of several voxels, spectra must be normalized after the application of the Fourier transform. The first few steps are mainly devoted to the removal of unwanted information which data contains, such as noise or baseline [2].

Phase correction

FID contains a real and imaginary part. In an ideal case there is no phase shift between the real and imaginary part. However, in real world experiments a phase correction must be applied to a time domain FID to obtain the proper spectrum.

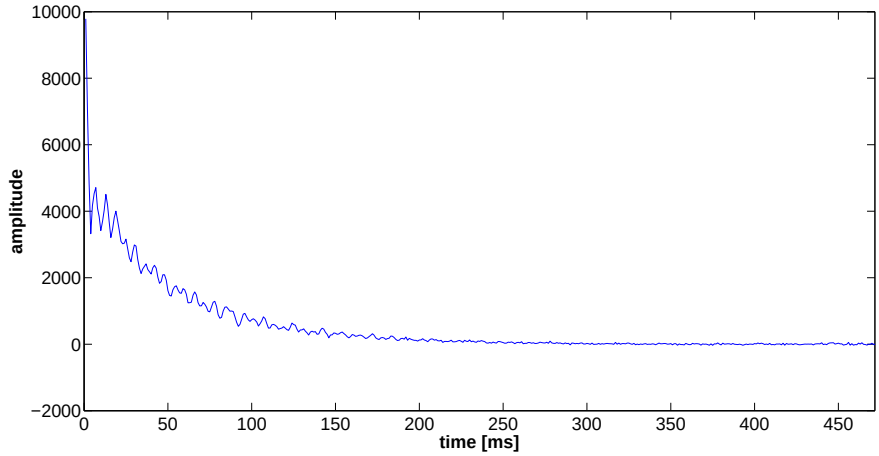

Figure 3. Exemplary FID signal obtained by proposed software. Data point index means position of point in data vector.

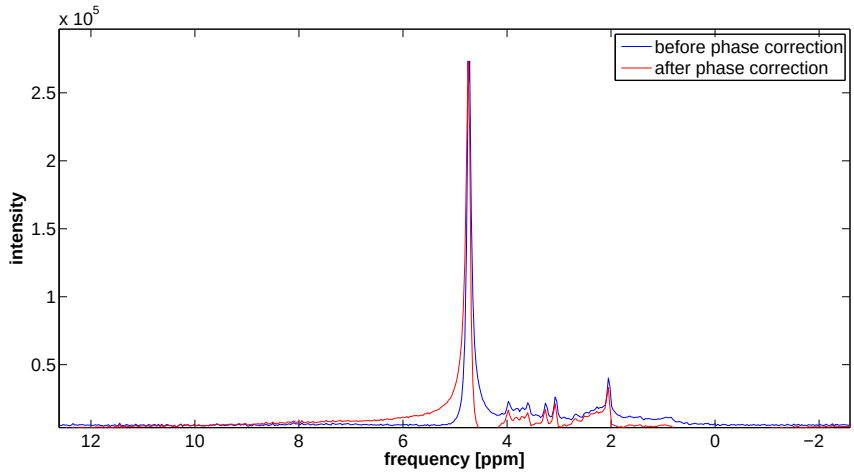

Figure 4. Signals: before (blue) and after(red) auto phase correction. One may notice that peaks on the right hand side of reference peak are better separated from each other.

An automatic phase correction is proposed. The iterative procedure starts with the initial value of phase correction and the corrected spectrum is calculated following the method described by Hornak [10]. The whole procedure is repeated for a defined set of correction angles. The quality criterion, that all meaningful peaks are present on positive side of OX axis, allows for choosing the best angle.

Noise filtering

The NMR spectrum contains noise that results from the process of magnetic resonance itself and imperfect measurement methodology. In order to filter out the noise different filters were examined and finally, the Savitzky-Golay [4] filter was chosen. The results obtained with this filter were the most accurate according to the following spectrum quality criterion: (i) peak/signal amplitude cannot be attenuated; (ii) signal-to-noise-ratio (SNR) should be at the level of 5% [1].

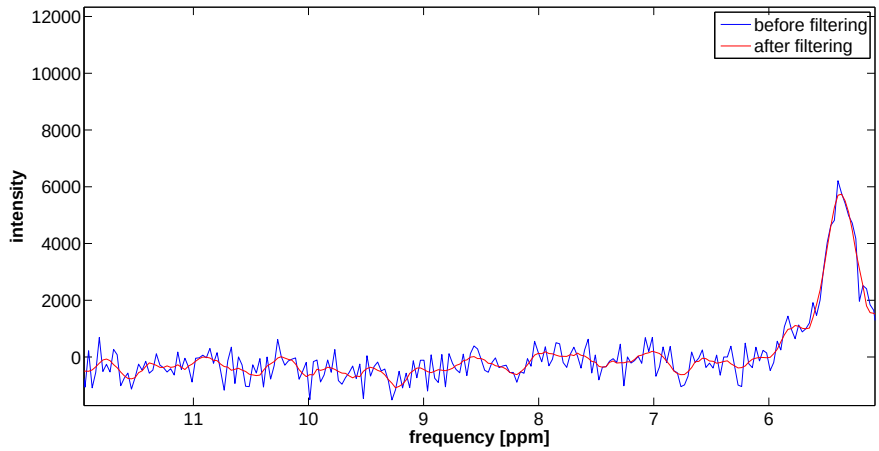

Figure 5. Signals: before (blue) and after (red) filtering procedure.

Baseline removal

NMR data usually contains a baseline resulting from signal of small amplitudes received from aminoacids. Number of techniques allowing to cope the problem have been developed. The most popular base on nonparametric baseline modeling. The alternative ones, parametric methods, model the baseline as a weighted exponentially decayed function. All these methods were examined and the median method was finally chosen. This method is simple, widely used and the results are acceptable. In median method it is necessary to create a moving window (a mask of size n). The baseline level is defined as equal to the median value of the signal inside the mask per each data point. Obtained in this way baseline is simply subtracted from the original signal. The result is presented in Fig. 6.

2.3. Water peak removal

Chemical structure of the human body causes that the leading metabolite which is able to be detected in 1H NMR, is water. According to medical investigation, the human body contains up to 50 (adults)or 75 (children)% of water. As a separate organ, the brain

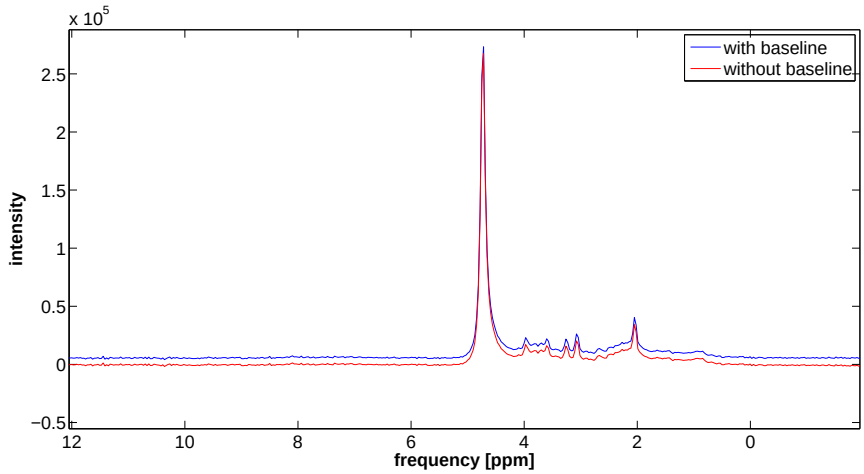

Figure 6. Signals: before(blue) and after (red) baseline subtraction.

contains 70% of water. It makes the peak of the water signal 100 times greater than the other metabolites. To ensure metabolite detection, the water signal must be suppressed from the overall spectrum. It is done during an automatic procedure at the beginning of the scan. However, we must take into account that presence of the water peak also brings some advantages. Knowing that water gives the highest signal, one may use it as a reference. It is possible to determine the chemical shift for water and then, using spectral resolution, chemical shifts for the other metabolites/peaks can be determined.

Literature study brings a lot of methods devoted to the removal of the water peak. The most popular is Hankel-Lanczos singular value decomposition [16] (abr. HLSVD). This method is used in many software applications. It requires modeling of the signal as sum of damped sinusoids. Using such a frequency selective filter, one may use it to select frequencies to be removed (e.g. water).

It is very important to control results during the process of water suppression. Sometimes an improper choice of parameters may lead to large baseline distortion. All the mentioned methods were examined. Because of the demand that the system should analyze the entire spectrum, another method was proposed. The water peak is determined (according to its known position) and in the next step the peak is replaced by a generated Gaussian noise with variance estimated with the use of the whole spectrum. The result is presented in Fig. 7.

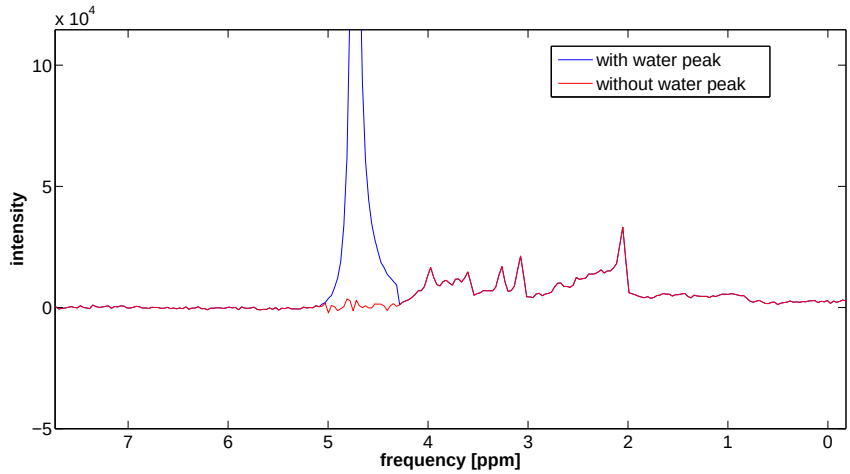

Figure 7. Signals: with suppressed water (red) and without suppressed water (blue). One may notice that signals differ only in water peak area.

3. Gaussian mixture model

Calculation of metabolite amount

NMR data form a set of peaks characterized with their width, height and position. In order to interpret these characteristics, a few important steps must be performed. The most important is calculation of metabolite amount. It is necessary to determine the area under a specified peak or group of peaks. There are some possible methods to be applied:

- *Numerical integration* [3]. It is the most traditional way of determining peak area in the frequency domain. The requirement is to define integration limits (the beginning and end of the selected peak). One must be aware that only the total peak area corresponds to the real peak intensity/metabolite amount. Obviously, some of the area (which is still under resonance) is beyond upper and lower limit of the integration. This causes underestimation of the result. Integration gives proper results only if the peaks are well separated from each other. However, in most in vivo tests, the peaks have tendencies to overlap.

- *Peak fitting* [3]. In the peak fitting method all the signals (that may be used in diagnostic process) are selected. Estimations of resonance frequencies, line width and intensity are performed as well. A fitting algorithm uses least squares optimization. It is an iterative method that fits all the chosen peaks to a line-shape model. Thanks to that, the resulting spectrum resembles the original one as close as possible.

- *Peak fitting with prior knowledge* [3]. This method gives an opportunity to use knowledge about metabolites, which may be observed in 1H in vivo spectroscopy. It is worth to notice, that the prior knowledge method is the only one which allows to reduce degree of freedom with an increase of obtained model accuracy.

- *Advanced prior knowledge with usage of metabolites sets* [3]. The main assumption of this method is the fact, that there exists a limited number of already studied and investigated metabolites that may be visible in 1H human brain spectroscopy. Spectra are analyzed as a linear combination of results recorded for individual metabolites.

Idea of Gaussian mixture

Following the idea of simultaneous modeling of the whole signal instead of every distinct peak, Gaussian mixture model (GMM) is proposed. GMM is a probabilistic model in which densities are estimated by a mixture distribution. It is developed from the mixture of k Gaussian components which are independent Gaussian distributions. Every component may be described by a probability density function [15].

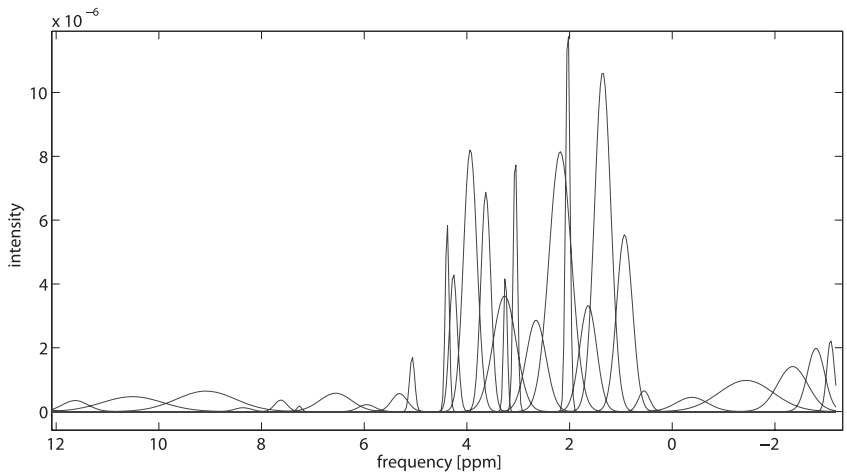

Figure 8. Exemplary result obtained for preprocessed data. Different components presented with different color.

After computation of k components, it is possible to construct a GMM as a mixture of every distinct component.

Expectation-maximization algorithm

The expectation maximization (EM) algorithm is a method that is widely used in recursive estimation of unknown parameters in a case of incomplete information [14].

We define a mixture distribution:

$$f^{mix} = (x_1, \alpha_1, .., \alpha_k, p_1, .., p_k) = \sum_{K=1}^{K} \alpha_k f_k(x_k, p_k) \tag{1}$$

where p are probability distributions and α_k are weights. Important assumption that:

$$\sum \alpha = 1 \tag{2}$$

must be satisfied. EM algorithm may be described as two following steps.

E- expectation step

In the E step it is necessary to assume initial values for parameters of probability distributions: α, μ, σ. Using assumed parameters, one may write the probability distributions as follows:

$$p(k|x_n, p^{old}) = \frac{\alpha_k^{old} \exp[\frac{-(x_n - \sigma_k^{old})^2}{2(\sigma_k^{old})^2}]}{\sum_{K=1}^{K} \alpha_k^{old} \exp[\frac{-(x_n - \sigma_k^{old})^2}{2(\sigma_k^{old})^2}]} \tag{3}$$

Having the probability distributions, it is possible to check whether the logarithmic likelihood condition is satisfied:

$$\ln[f(x^e, p)] = \sum_{n=1}^{N} \ln \alpha_k + \sum_{n=1}^{N} f_k(x_k, p_k) \tag{4}$$

This condition may be treated as satisfied if the difference between two iterations of the EM algorithm is significantly small. It is important to notice that such difference should be properly defined. During the first iteration it is useless to terminate calculations because the result is certainly incorrect. To perform the second iteration, one may predefine the 1st difference between logarithmic likelihood, as a value greater than critical difference (needed to finish the computations).

M-maximization step

In the second step of the EM algorithm new values of parameters are calculated as follows:

$$\alpha_k^{new} = \frac{\sum_{n=1}^{N} p(k|x_n, p_{old})}{N} \tag{5}$$

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} x_n p(k|x_n, p_{old})}{\sum_{n=1}^{N} p(k|x_n, p_{old})} \tag{6}$$

$$(\sigma_k^{new})^2 = \frac{\sum_{n=1}^{N} (x_n - \mu_k^{new})^2 p(k|x_n, p_{old})}{\sum_{n=1}^{N} p(k|x_n, p_{old})} \tag{7}$$

Iterations are repeated until a successful attempt of obtaining a significantly small difference between likelihood functions are obtained. To use EM in a case of NMR modeling, a few modifications were proposed. Instead of number of components, their lower and upper limits are predefined. Computations are performed, until the optimal solution is obtained. To assure its convergence, the algorithm must be run about 150-200 times (for different initial conditions). When optimal solution for given number of Gaussian components is obtained, the value of the Bayesian Information Criterion (abr. BIC) is calculated [6] as follows:

$$BIC = -2\ln(L) + (3k-1)\ln(n) \tag{8}$$

where: L is maximized value of the likelihood, k is number of GMM components and n is length of data string.

An optimal solution (for the number of components) is chosen for the largest BIC value. As a result the set of probability distribution parameters is obtained. The length of every parameters vector corresponds to the optimal number of Gaussian components.

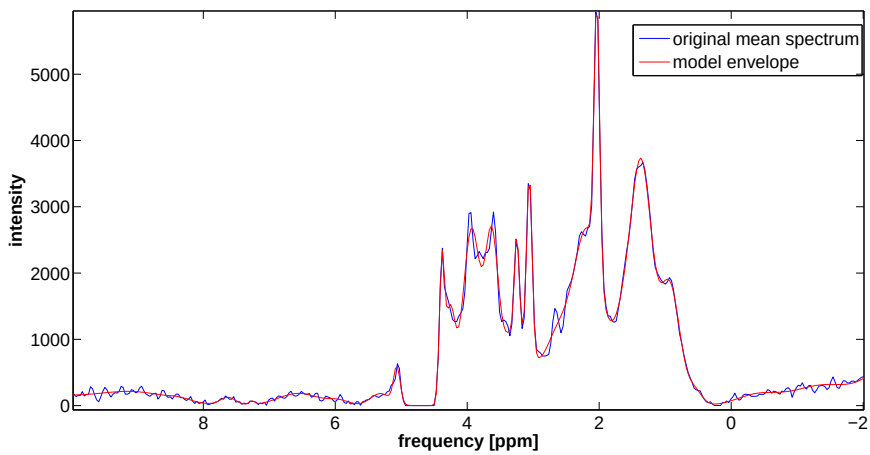

Figure 9. Signal and obtained GM Model presented on one plot. Model (red) and original signal (blue). Model was rescaled in order to envelope the signal.

To investigate whether the number of components was chosen correctly, the results of EM are to be compared with the simple peak detection routine. Numerical results are presented in Tab. 1.

As one may notice the number of detected peaks and number of Gaussian components are similar which leads to the conclusion, that the number of components obtained by means of EM is proper.

Using the obtained GMM model one can calculate the amount of a specific metabolite in every voxel of the tested area by a weighted convolution of Gaussian component

Table 8. Comparison between number of significant peaks and number of components detected by EM algorithm. Threshold value is the level that all valid peaks must equal or exceed.

Sample	Threshold value	Peak detection	Number of EM components
1	0.5	69	41
	1	44	
	2	18	
	3	9	
2	0.5	72	43
	1	46	
	2	17	
	3	9	
3	0.5	65	40
	1	47	
	2	20	
	3	8	

and normalized NMR signal. After calculating the metabolite amount per one voxel results are reordered into original voxel placement. Because of the discretization of the tested area, the obtained heatmap is not smooth enough. In order to approximate the signal level in between the voxel centers, cubic spline method was applied [7]. Comparison of discrete and smoothed heatmaps is presented in Fig. 10.

4. Exemplary results

4.1. Comparison with existing solutions

At the stage of preprocessing, it is possible to compare results obtained with the use of the proposed methodology with ones given by software already available on the market. Four applications were examined: Tarquin, AQSES, iNMR and jMRUI. Most of them allows only to compare results graphically as there is no possibility to export results in a numerical form. Results of the comparison are presented in Fig. 11. The small differences observed are probably caused by numerical simplification

4.2. Results obtained by ‘GMM-NMR’

To support cancer diagnosis and treatment planning, an evaluation of tumor size was performed with the use of obtained metabolite maps. First, a metabolite heatmap was

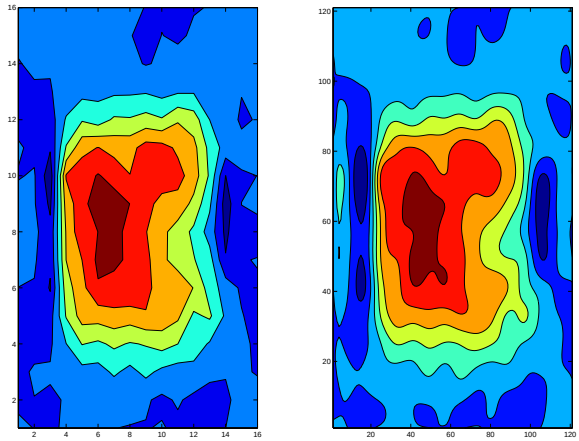

Figure 10. The comparison of discrete (left) and smoothed (right) heatmaps.

blended with original MR images. Choline metabolite concentration was chosen as the marker because of its correlation with a process of cell death. Results are presented in Fig. 12-14. On the right pictures, the red circle roughly shows the tumor area.

5. Conclusion

The proposed solution may be used in estimation of tumor size as a tool in cancer diagnosis and treatment planning. Results were shown only for one metabolite, however in order to precisely estimate tumor size and location, it is recommended to look at the heatmaps of all tumor related metabolites. The analysis can be performed in fully automated mode, no user parameter definition is required. The system is stable and user-friendly. It allows for the importing of final results in the form of graphical and text files.

The main drawback of the existing solution is that the estimated Gaussian components' locations are sensitive toward the sample quality. It might happen that the mean values are slightly different from original peak placement. It is possible to modify the EM algorithm in such a way, that some of the mean values could be predefined before the algorithm runs. Such a model may be then applied to every single voxel that would lead to a very accurate solution. Such a solution is planned for implementation in near future.

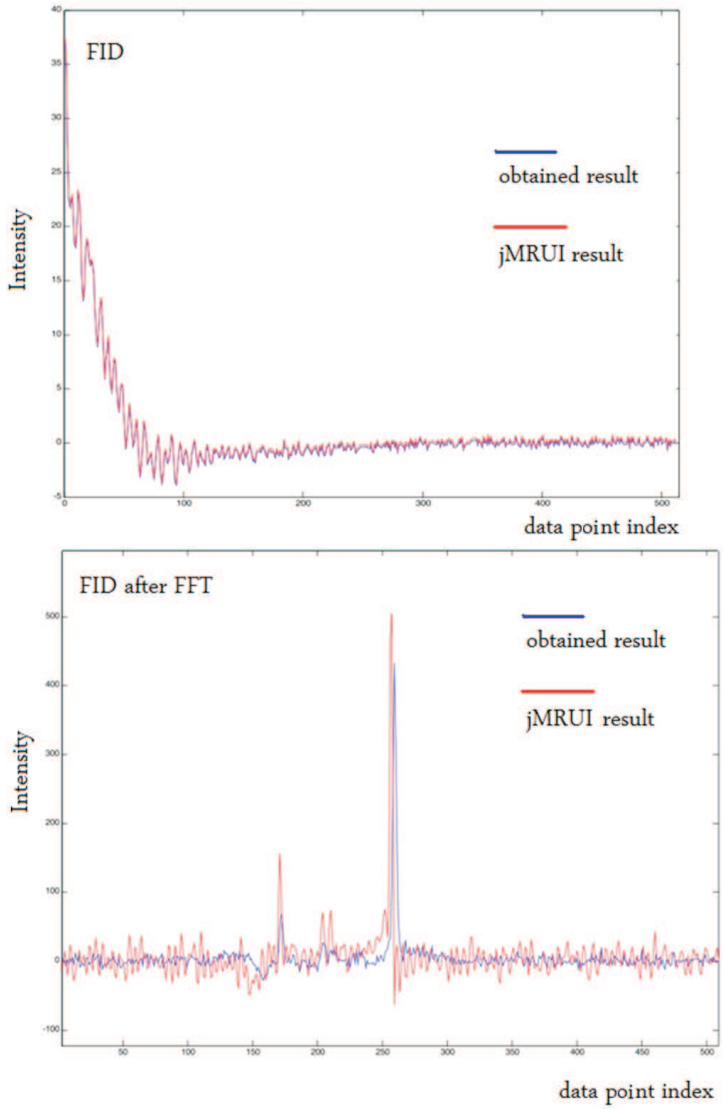

Figure 11. Result of comparison between obtained solution and already implemented and existing g on the market proposed by jMRUI. Data point index means position of point in data vector.

References

[1] R.A. DE GRAFF: In vivo NMR spectroscopy. John Wiley and Sons Ltd., 2007.

[2] J.F. JANSEN, H.W. BACKES, N. KLAAS and M.E. KOOI: 1H MR spectroscopy of the brain: Absolute quantification of metabolites. *Radiology*, **240** (2007), 318-333.

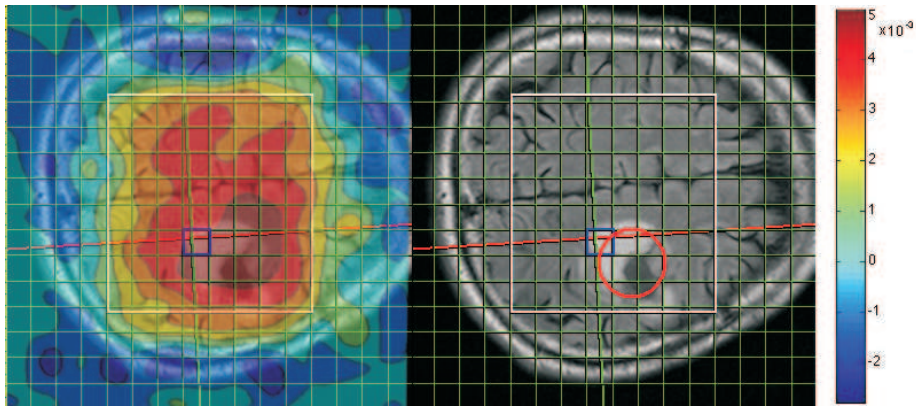

Figure 12. Exemplary result case 1.

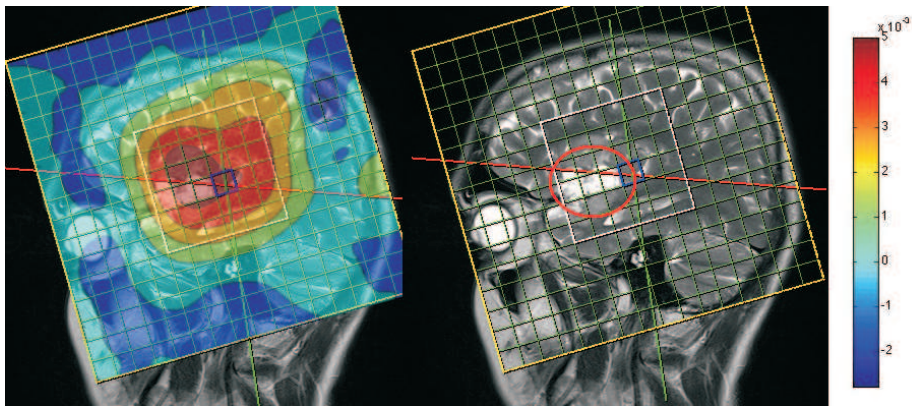

Figure 13. Exemplary result case 2.

[3] J. KEELER: Understanding NMR spectroscopy. John Wiley and Sons Ltd., 2005.

[4] A. SAVITZKY and M.J.E. GOLAY: Smoothing and differentiation of data by Ssimplified least squares procedures. *Analytical Chemistry*, **36**(8), (1964), 1627-1639.

[5] A. POLAŃSKI and M. KIMMEL: Bioinformatics. Springer Verlag Berlin Heidelberg, 2007.

[6] P.E. MILLAR: Using the Bayesian information criterion to judge models and statistical significance. North American Stata Users' Group Meetings, 2006.

[7] S. MC KINLEY and M. LEVINE: Cubic spline interpolation. Math 45: Linear Algebra, 1992.

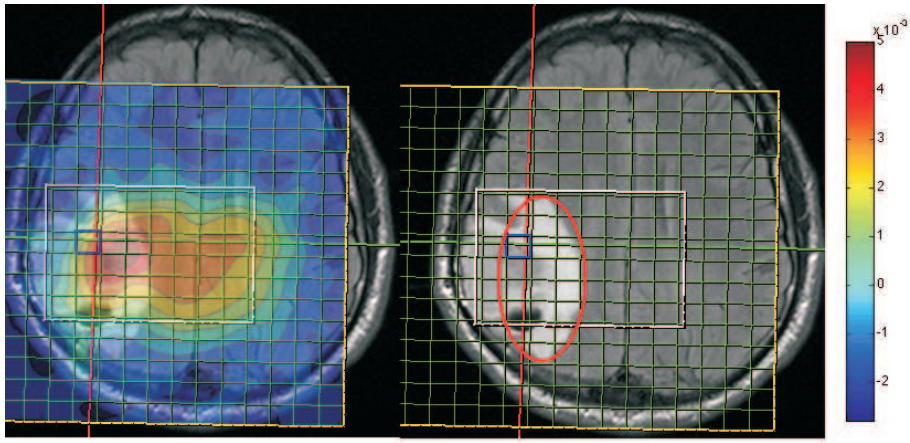

Figure 14. Exemplary result case 3.

[8] M. PIETROWSKA, L. MARCZAK, J. POLAŃSKA, E. NOWICKA, K. BEHRENT, R. TARNAWSKI, M. STOBIECKI, A. POLAŃSKI and P. WIDLAK: Optimizing of MALDI-ToF-based low-molecular-weight serum proteome pattern analysis in detection of breast cancer patients; the effect of albumin removal on classification performance. *Neoplasma*, **56**(6), (2010), 537-44.

[9] webpage: http://tarquin.sourceforge.net/index.php, Tarquin project, acces 6.12.2010.

[10] J. GONG and J.P. HORNAK: A fast T1 algorithm. *Magn. Reson. Imaging*, **10** (1992), 623-626.

[11] D. SHAW: Fourier transform NMR spectroscopy. Elsevier, NY, 1976.

[12] J-B. POULLET: An automated quantitation of short echo time MRS spectra in an open source software environment: *AQSES NMR in Biomedicine*, **20**(5), (2007), 493-504.

[13] D. STEFAN, F. DI CESARE, A. ANDRASESCU, E. POPA, A. LAZARIEV, E. VESCOVO, O. STRBAK, S. WILLIAMS, Z. STARCUK, M. CABANAS, D. VAN ORMONDT and D. GRAVERON-DEMILLY: Quantitation of magnetic resonance spectroscopy signals: the jMRUI software package. *Measurement Science and Technology*, **20**(104035), (2009).

[14] A.P. DEMPSTER, M.N. LAIRD and D.B. RUBIN: Maximum likelihood from incomplete data wia the EM algorithm. *Journal of the Royal Statistical Society B*, **39**(1), (1977), 1-22.

[15] G. MCLACHLAN (ED.): Mixture models. Marcel Dekker, NY, 1988.

[16] E. CABANES, S. CONFORT-GOUNY, Y. LE FUR, G. SIMOND and P.J. COZZONE: Optimization of residual water signal removal by HLSVD on simulated short echo time proton MR spectra of the human brain. *J.of Magnetic Resonance*, **150**(2), (2001).