MATHEMATICAL SIMULATION OF MASS SPECTRUM

L.Beránek[1], J.Knížek[2], Z. Pulpán[3], M. Hubálek[4], V. Novák[1]

[1]University of South Bohemia, Ceske Budejovice, [2]Charles University, Hradec Kralove, [3]University of Hradec Kralove, Hradec Kralove, [4]University of Defense, Brno

Abstract

The contribution presents shortly simulation of mass spectrum. This was necessary for debugging and testing of the mathematical algorithms for the processing of data from mass spectroscopy

1 Introduction

Mass spectra represent very valuable information source which can be used at solving research and diagnostics problems in biology and experimental medicine. Many studies deal (for example [1], [2], [3], [4]) with the problem of statistic evaluation of mass spectra for the purposes of biology research. Long line of works uses these statistical procedures to the evaluation of mass spectra within the scope of biology research.

Mass spectrum can be described as a dependence of relative ion percentage intensity on its effective mass. Relative ion percentage intensity is the intensity of ion related to the intensity of maximum ion in a given spectrum. It is denoted in literature by the character $I[\%]$. Effective mass is the ratio of ion mass and its charge $m/z\,[1]$. This ratio is characteristic for every particle. Mass spectrum can serve to perform an exact identification of a given particle under the condition of sufficient resolution of instrument and its precise calibration.

2 Data

Mass spectrometry generates (primary) data: single shots of mass spectral values $y(x_t)$ for every x_t value, $t = 1,\ldots,T$, of abscise vector, i.e. for $\boldsymbol{x} = (x_1,\ldots,x_T)$, where T is the number of abscises (or measured places), $T \approx 10^4$.
The data structure of an mass spectrum can be mathematically written in the form:

$$\left.\begin{matrix} y_1 & \cdots & y_T \\ x_1 & \cdots & x_T \end{matrix}\right\}. \tag{1}$$

The less brief notation, that could be also used, is $y_t \equiv y(x_t)$, where $t = 1,\ldots,T$. The spectrum $y(x)$ is normalized so that the maximum of the highest „peak“ gets the value $y_{\max}(x) = 100\%$. The value of independent variable x moves approximately within the interval $\langle 0\ ;\ 1.2\cdot 10^5\rangle = \langle l\,;u\rangle$. The real 1-shot mass spectrum is displayed on the fig. 1.

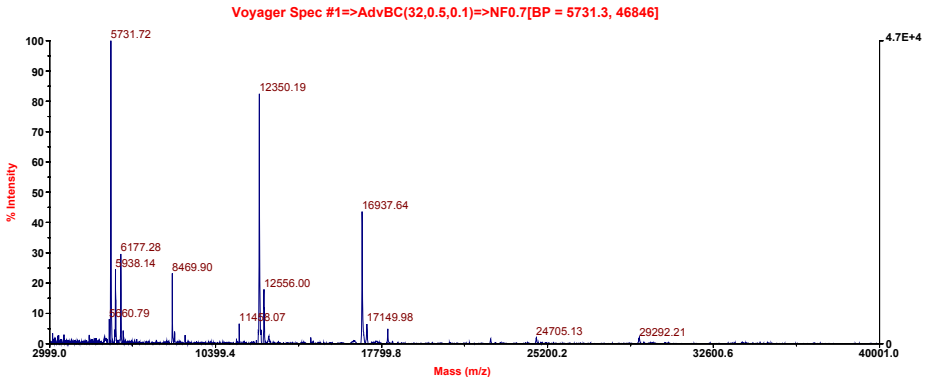

Figure 1: mass spectrum (Voyager - DE STR, Voyager instrument control panel version 5.1, Data explorer version 4.5).

3 Mathematical simulation of mass spectrum

Let us suppose that the random sample is an experimental representation of a certain random quantity $Y(x)$ which represents considered spectrum. Argument x gathers only the known values $x \in \{x_1,\ldots,x_T\}$. The random quantity $Y(x)$ can be expressed in the form

$$Y(x) = y(x) + e(x), \quad x \in \{x_1,\ldots,x_T\}, \tag{2}$$

where $e(x)$ consists of residual error after the processing of rough spectra, of random disturbances which represent e.g. biological variability, and of laboratory error of experiments e.g. in molecular biology etc. $y(x)$ is the exact value.

3.1 One peak simulation

A peak is defined by its coordinates x and y of peak maximum, i.e. by the numbers x_0 and y_0, and by the ratio of height and width of peak R at half its height.

The modified functional dependence of normal distribution probability density function in the form

$$f(x) = \frac{K}{\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in \langle l;u \rangle \tag{3}$$

can be used for peak simulation. To achieve the correct shape of the peak whose right side "tail" is always visibly higher then the left side "tail", the peak is modeled with the help of normal distribution probability function so that the left side peak is expressed by the equation (3), and for the right side the equation (3) is modified so that σ is multiplied by the empirical coefficient ζ. Hence the left peak side:

$$f_L(x) = \frac{K_L}{\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in (l;\mu\rangle, \tag{4}$$

right peak side:

$$f_R(x) = \frac{K_R}{\zeta\sigma} e^{-\frac{(x-\mu)^2}{2(\zeta\sigma)^2}}, \qquad x \in (\mu; u\rangle\,. \tag{5}$$

Further it has to hold that the ration of peak height and width in its half height is constant:

$$y_0 /(z_R - z_L) = R\,, \tag{6}$$

where $z_L < z_R$ are related peak x-coordinates in the half of its height. Thus, it applies $f_L(x = z_L) = f_R(x = z_R) = y_0 / 2$, whereas $y_0 = f_{L,\max}(x_0) = f_{R,\max}(x_0)$.

3.1.1 Derivation of computational formulas for z_L and z_R

It can be easily derived from the relation (3) (or (4) and (5)) that the coordinate x of peak maximum is identical to the position parameter μ: $x_0 \equiv \mu$, thus

$$f_L(x_0) = y_0 = K_L / \sigma \qquad \text{a} \qquad f_R(x_0) = y_0 = K_R /(\zeta\sigma)\,. \tag{7}$$

E.g. for the left peak side it is valid

$$\frac{y_0}{2} = f(z_L) \qquad \Rightarrow \qquad \frac{K_L}{2\sigma} = \frac{K_L}{\sigma} e^{-\frac{(z_L-\mu)^2}{2\sigma^2}}\,. \tag{8}$$

After the arrangement:

$$\frac{1}{2} = e^{-\frac{(z_L-\mu)^2}{2\sigma^2}}\,. \tag{9}$$

After arrangement of (9) we receive the quadratic equation

$$z_L^2 - 2\mu z_L + \mu^2 - 2\sigma^2 \ln 2 = 0\,, \tag{10}$$

and after its solving the result has form $z_L = z_L(\sigma) = \mu - \sigma\sqrt{\ln 4}$. Analogously, for the right peak side $z_R = z_R(\zeta\sigma) = \mu + \zeta\sigma\sqrt{\ln 4}$, where ζ is the coefficient which enables to modify the shape of the right peak „tail“, $\zeta \approx 5$.

Explicit relations for the calculation of constants K_L or K_R can be then expressed by the help of relations (7) (providing σ is known, see below): $K_L = y_0\,\sigma$ and $K_R = y_0\,\zeta\,\sigma$.

3.1.2 The calculation of σ parameter

To calculate parameter σ it is necessary to solve transcendental equation (e.g. by the method regula falsi)

$$\frac{y_0}{z_U(\zeta\sigma) - z_L(\sigma)} - R = 0 \tag{11}$$

for the unknown variable σ. The value of $z_L(\sigma)$ and $z_U(\zeta\sigma)$ can be calculated by means of relations from the last section. The iteration process is to be repeated so long until the value σ_r does not change in two successive iterations with an accuracy of required number of significant figures d:

$$\frac{|\sigma_r - \sigma_{r-1}|}{|\sigma_r|} \le \frac{1}{2} 10^{-d}, \tag{12}$$

where r is the order number of the final iteration. The solution should always find itself in the interval $\sigma_r \in \langle \sigma_L; \sigma_R \rangle$, where $\sigma_L = 10^{-13}$ and $\sigma_U = 10^2$, for $x_0 \equiv \mu \in \langle l; u \rangle$, $y_0 \in \langle 1\%; 100\% \rangle$ and $R \approx 10^3$.

It is necessary to solve by means of regula falsi the modified transcendental equation with regard to big order difference between σ_L and σ_R:

$$\frac{y_0}{z_U(10^{\sigma' + \log_{10}\zeta}) - z_L(10^{\sigma'})} - R = 0 \tag{13}$$

for unknown $\sigma'_r = \log_{10}\sigma_r$ and correspondent bounds $\sigma'_L = -13$ and $\sigma'_R = 2$.

The result of these calculations is demonstrated in the picture 2.

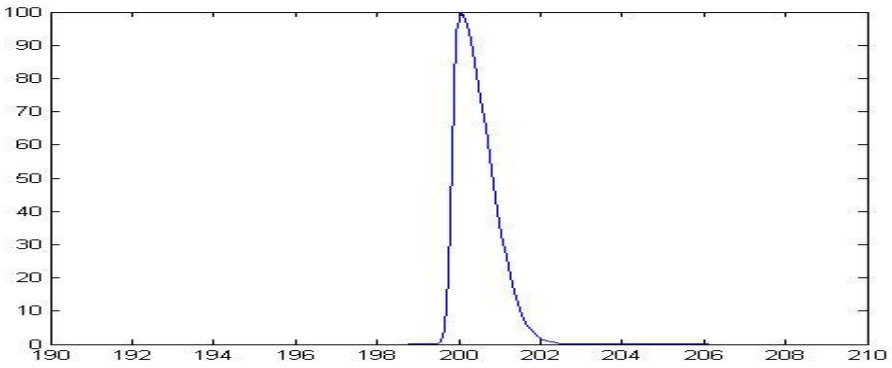

Figure 2: One peak simulation

3.2 Spectrum simulation

Ideal (un-noised by random disturbances) dependence of *m*-peaks simulated spectrum can be written in the explicit relation

$$y(x) = \sum_{j=1}^{j=m} f(x, x_{0,j}, y_{0,j}, \mu_j, \sigma_j, K_{L,j}, K_{U,j}, \zeta), \tag{14}$$

where $f(.)$ are functions of particular peaks from § 3.1. X-coordinates of maximums $x_{0,j}$, $j = 1, \ldots, m$, of all peaks, i.e. of proteins (and their fragments), are in all n shots of spectra

identical. They are as though natural constants. Their values are generated with the help of generators of random number of rectangular distribution in the interval of all measured range of x spectra coordinates, i.e. in the interval $\langle 0\,;1.2\cdot 10^5\rangle = \langle l\,;u\rangle$. The heights of all peaks $y_{0,j}$, $j=1,\ldots,m$ are generated with the help of generator of random number of rectangular distribution in the interval $(1\%;100\%)$.

The heights of all peaks $y_{0,j}, j=1,\ldots,m$, represent in the framework of one n-shots mass spectrum as though constants. But at simulation of data (represented by more n-shots mass spectra) these heights of all peaks $y_{0,j}$, $j=1,\ldots,m$, are similarly random quantities noised by random disturbances which represent e.g. biological variability, laboratory error of experiments e.g. in molecular biology etc.

Ideal spectrum calculated according the described steps in various scales is demonstrated in the pictures 2 and 3.

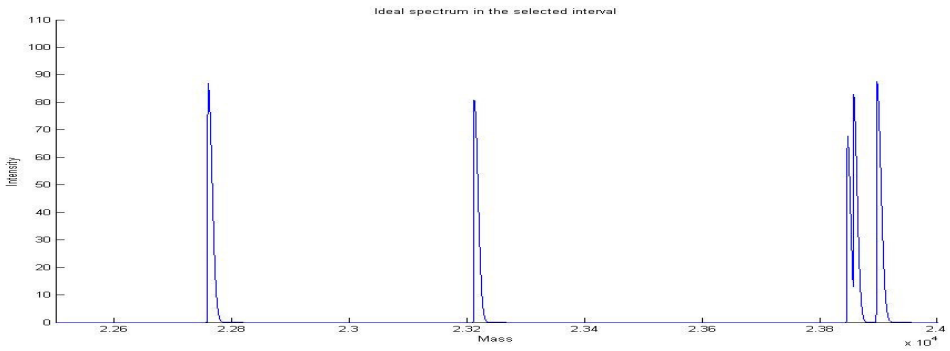

Figure 3: Ideal spectrum simulation in the selected interval

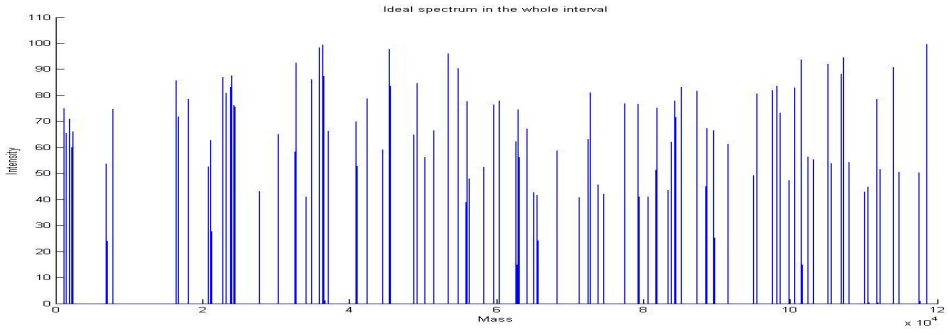

Figure 4: Ideal spectrum simulation in the whole interval

4 Conclusion

The purpose of our future work is to design the new methodology of mathematical-statistical and fuzzy-logical identification and decision making in the domain of protein biomarkers from mass spectra. The described simulation of mass spectrum is necessary for debugging and testing of the mathematical algorithms for the processing of data from mass spectroscopy. For these purposes the Matlab environment is very proper tool.

References

[1] J. Knížek, Z. Pulpan, M. Hubalek, L. Beranek, P. Pokorny. *Stochastic model of mass spectrum ramdom distribution and its simulation*, to be published in Journal of Mass Specktrometry

[2] G. Ball, S. Mian, F. Holding, R.O. Allibone, J. Lowe, S. Ali, G. Li, S. McCardle, I.O. Ellis, C. Creaser, R.C. Rees *An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers.* Bioinformatics. 2002, Mar; 18(3):145-164.

[3] E.P. Diamandis. *Mass Spectrometry as a Diagnostic and a Cancer Biomarker Discovery Too.*, Molecular and Cellular Proteomics, 2004, 3:127-114.

[4] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, Q.T. Le. *Sample classification from protein mass spectrometry by "peak probability contrasts".* Bioinformatics - Bioinformatics Advance Access, Oxford University Press, 2004, 1-10.

Author1
Contact information
Department of Computer Science, Faculty of Education in Ceske Budejovice, University of South Bohemia, Jeronymova 10, 371 15 Ceske Budejovice, Czech Republic, e-mail: beranek@pf.jcu.cz
Author2
Contact information
Department of Medical Biophysics, Faculty of Medicine in Hradec Kralove, Charles University in Prague, Simkova 870, 500 38 Hradec Kralove, Czech Republic
Author3
Contact information
Department of Mathematics, Faculty of Education, University of Hradec Kralove, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic
Author4
Contact information
Institute of Molecular Pathology, Faculty of Military Health Sciences, University of Defense in Brno, Trebeska 1575, 50001 Hradec Kralove, Czech Republic
Author5
Contact information
Department of Computer Science, Faculty of Education in Ceske Budejovice, University of South Bohemia, Jeronymova 10, 371 15 Ceske Budejovice, Czech Republic