Available online at www.sciencedirect.com
ScienceDirect
International Journal of Mass Spectrometry 260 (2007) 20–30
International Journal of Mass Spectrometry
www.elsevier.com/locate/ijms

Automated reduction and interpretation of multidimensional mass spectra for analysis of complex peptide mixtures

Anna Gambin [a]*, Janusz Dutkowski [a], Jakub Karczmarski [b], Bogusław Kluge [a], Krzysztof Kowalczyk [a], Jerzy Ostrowski [b], Jarosław Poznański [c], Jerzy Tiuryn [a], Magda Bakun [c], Michał Dadlez [c,d]

[a] *Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland*
[b] *Department of Gastroenterology, Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, 02-781 Warsaw, Poland*
[c] *Institute of Biochemistry and Biophysics PAS, Pawińskiego 5A, 02-106 Warsaw, Poland*
[d] *Biology Department, Warsaw University, Miecznikowa 3, 02-096 Warsaw, Poland*

Received 13 January 2006; received in revised form 1 June 2006; accepted 1 June 2006
Available online 4 August 2006

Abstract

Here we develop a fully automated procedure for the analysis of liquid chromatography–mass spectrometry (LC–MS) datasets collected during the analysis of complex peptide mixtures. We present the underlying algorithm and outcomes of several experiments justifying its applicability. The novelty of our approach is to exploit the multidimensional character of the datasets. It is common knowledge that highly complex peptide mixtures can be analyzed by liquid chromatography coupled with mass spectrometry, but we are not aware of any existing automated MS spectra interpretation procedure designed to take into account the multidimensional character of the data. Our work fills this gap by providing an effective algorithm for this task, allowing for automated conversion of raw data to the list of masses of peptides.

Keywords: LC–MS; Deconvolution; Mass decomposition; Clustering; Biomarker selection

1. Introduction

Rapid development of mass spectrometric (MS) technology offers the possibility of performing exhaustive analysis of complex mixtures containing thousands of molecules in a single experiment. In a typical experiment a complex mixture of species (most often peptides) is separated using liquid chromatography coupled on-line with electrospray mass spectrometer. In result a single LC–MS experiment produces a 2D dataset in which each detected peptide is characterized by two coordinates—its molecular mass over charge (m/z) value and retention time value. Such datasets constitute a so called "peptide mass fingerprint" of the sample.

Peptide mass fingerprints (PMF) are of the interest because of their potential application as a completely new tool for medical diagnostics of diverse conditions. The most common starting material for obtaining PMF are human body fluids, mainly blood plasma or serum. Much work has already been invested into detection of molecular mass biomarkers for various pathologies and diagnostic procedures have been suggested [1,26,9,20–24,27,31,39].

Initially great hopes for early diagnostics of ovarian cancer were advocated by the inventors of SELDI technology [25]. However, their optimism has encountered strong criticism [6,7,10,12]. The criticism was addressed not against the idea of using PMF as a diagnostic tool but against a poor quality of data obtained with SELDI approach. SELDI approach assumes first a protein extraction step and next acquisition of a simple 1D MALDI mass spectrum. Inadequate quality of produced data is caused by:

(A) poor resolution of mass spectrum where individual peaks are not resolved, and cannot be identified,
(B) matrix interference leading to loss of data at least below 500 Da,
(C) poor reproducibility of protein extraction step.

* Corresponding author. Tel.: +48 22 55 44 577; fax: +48 22 55 44 400.
E-mail address: aniag@mimuw.edu.pl (A. Gambin).

1387-3806/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijms.2006.06.011

LC–MS approach, producing PMF of well resolved peptide signals is a possible alternative to SELDI. It allows to overcome the above problems at a cost of longer time of analysis (minutes as compared to seconds in SELDI). In addition LC–MS allows for identification of species present, because LC–MS–MS/MS sequencing of selected species is routine. The ability to identify the sequence of a given biomarker certainly increases its value.

PFM datasets resulting from different separation techniques are recently intensively analyzed [18,33]. PFM produces large datasets, their manual analysis is not feasible, whereas proper software tools have only started to appear. Such software tools are necessary to convert PMF datasets, sometimes containing more than 10^4 species, in an automated way, into a list of molecular masses along with their retention times and signal amplitudes. Several attempts to approach the above problem have been undertaken. Authors of [8] adopt the pattern recognition techniques to process the LC–MS dataset. Such an approach results in a very effective peak picking procedure but it does not allow for correct deconvolution (i.e., identification of isotopic envelopes and charge determination). In [37] a logical idea has been proposed, mainly, to explore the analogy between MS and microarray technology for transcriptomics. A new tool called SpecArray generates a peptide versus sample array from a set of LC–MS data. Unfortunately the software itself is not yet accessible. Therefore we could not compare that work with our approach.

The intrinsic multidimensional character of PMF data (2D for LC–MS and 3D for MudPIT LC–LC–MS experiment) is not properly appreciated in existing software. Our work attempts to include data multidimensionality at an early step of data processing. The advantage of our approach can be illustrated be comparison with work carried out in [3] where the dataset of similar format were analyzed. However, the 2D spectrum was processed line by line in a distributed environment. The time required by our method to analyze the whole sample is roughly of the same order as the time needed by the distributed method of [3] to analyze two lines of the input (i.e., about 1% of the whole dataset). This significant speed-up is achieved by treating the data as 2D: the algorithm groups peaks in 2D spectrum and processes each group independently. This approach does not require line-by-line analysis of the spectrum and is more effective.

Our approach requires a preprocessing phase in which the noise reduction and peak picking is performed on PMF dataset. We have decided to use existing software—namely NMRPipe [11] tool and XCMS package [4]. The advantage of using NMRPipe comes from the fact that it has been tested by many groups for many years in NMR based protein structural studies. Since the numerical character of NMR data is the same as MS data NMRPipe has proven to be extremely effective after proper data format conversion procedures allowed MS data to be processed by NMRPipe.

Main steps of our PMF dataset analysis include: (1) noise filtration, (2) clustering into isotopic envelopes, (3) automated charge determination, (4) calculation of deconvolution significance and (5) monoisotopic mass calculation. Crucial postprocessing step includes aligning masses across samples. To this aim the retention times should be appropriately normalized. During all phases of processing an efficient visualization tool allows for manual inspection of PMF datasets. Sparky [13] tool is used for this purpose.

The noise filtration performed during preprocessing phase has to be refined at the beginning of our algorithm to eliminate peaks corresponding to spurious signals (Step 1).

The main challenge in the automated analysis of MS spectrum is to determine ion charge and to group isotopic peaks coming from the same ion. The main idea behind our approach here is to improve rough clustering of peaks done in Step 2 by exploring the raw dataset during the next step.

For automated charge determination, Zhang and Marshall [38] proposed ZScore algorithm, which uses a scoring scheme to assign charge for ions. The other method described by Senko at al. [29] is based on counting the Patterson and Fourier routines for selected areas of spectrum and then multiplying the results. For deconvolution, Senko et al. [30] proposed the "averagine" method, based on fitting an isotopic distribution from spectrum to theoretical distribution. This method is appropriate for large molecules (10–20 kDa). Horn et al. [16] proposed the algorithm THRASH based on both Patterson and Fourier routines and "averagine" method. For small peptides the problem of grouping isotopic peaks is relatively easy because the monoisotopic peak is usually the highest or the second highest. In our method we incorporate the ideas from [29] and [16] and adopt them to a 2D setting (Step 3).

For estimating the deconvolution significance the "averagine" model could not be applied for small molecules (up to 5000 Da). Instead, we propose our original method, which tests several models and chooses one that approximates well the distribution of abundance of different isotopic ions in the spectrum (Step 4). The quality of fit to the model corresponds to significance of a given isotopic envelope.

The monoisotopic mass of the peptide is assuming charge value which yields the most significant deconvolution (determined in the previous step). The amplitude of the signal is calculated as the sum of volumes of all peaks from its isotopic envelope (Step 5).

In the next section we describe the dataset, sample preparation and computational preprocessing. Then we describe our algorithm, its main features and its complexity. The accuracy, sensitivity and effectivity of the algorithm is analyzed and the preliminary results for classification and biomarker search are presented in Section 4. The directions for further work are sketched in Section 5.

2. Preprocessing

2.1. Blood plasma (serum) peptidome extraction and analysis

Blood samples were collected from patients and sex- and age-matched healthy controls. Blood plasma samples were collected in tubes with K3 E EDTA K3 (greiner bio-one cat. no. 455036) then centrifuged at 2800 × *g* 15 min at 4 °C. For serum collection EDTA was omitted. Obtained plasma (serum) was aliquoted in 200 μl portions, frozen in liquid nitrogen and stored at −70 °C for further use.

For analysis, plasma (serum) aliquots were centrifuged through a 5 kDa cutoff filtration membrane (Millipore Ultrafree-MC) in the presence of 20% acetonitrile as a chaotropic agent. Membrane was thoroughly washed prior to use. To the filtrate the internal standard was added. A HPLC purified peptide (220 pg) obtained from tryptic digest of lysozyme (FESNFNTQATNR, mol. mass: 1428.65 Da, $m/z = 714.82^{+2}$) was used as an internal standard.

After filtration plasma peptidomic fraction was subject to nano-HPLC coupled to MS measurements. PepMap columns (LC Packings) were selected as optimum for peptide mixture separations. Sample was first acidified by 0.1% trifuoroacetic acid and loaded on a pre-column. Afterwards the solvent is exchanged by 0.1% formic acid and the sample was transferred using an acetonitrile gradient 0–15% AcN in 100 min to a 75 μM nanoPepMap column coupled on-line to Q-ToF (Waters) or LTQ FTICR (Thermo) electrospray mass spectrometer working either in the regime of mass survey scan (for peptide mapping and relative quantitation) or in the regime of data dependent MS to MS/MS switch (for peptide identification). Sample LC–MS–MS/MS analysis is fully automated and multiple runs do not require any operator's intervention. AcN gradient was worked out experimentally for optimum separation of blood plasma (serum) peptides.

2.2. Data preprocessing: NMRPipe

Raw datafiles were subjected to data format conversion and analyzed by NMRPipe [11]. We have used NMRPipe to smooth out the data and improve the SNR (signal-to-noise ratio). The crucial task done by NMRPipe is 3D peak picking, i.e., searching for peaks (local maxima) in the spectrum. The outcome of this stage of processing is the list of coordinates of peaks in spectrum with the signal amplitude and the peak volume for each peak.

2.3. Data preprocessing: XCMS package

Results obtained with NMRPipe were compared with recently proposed XCMS R-package [4]. It provides procedures for filtration and peak picking as well as matching peaks across samples. It also performs the correction of retention time. XCMS package is included in the Bioconductor open source software project and can be downloaded from http://www.bioconductor.org. We have compared the efficiency and sensitivity of these two approaches, i.e., XCMS and NMRpipe (data not shown). Our observations are in favor of the NMR tool because of the more flexible visualization functions and the suitability for the high-throughput processing. Further work is based on peak lists generated with NMRPipe tool (Fig. 1).

Input: RL value threshold, *max_charge*
Output: list of peptides
Step 1: noise filtration: calculate the cutoff threshold
Step 2: isotopic clusters identification
  foreach isotopic cluster
Step 3: suggest the charge using Patterson–Fourier transform
Step 4: for suggested monoisotopic mass calculate the possible distributions of isotopic species (i.e., the **isotopic model**) in the cluster and estimate the significance of the cluster (figure-of-merit (FOM)) from Eq. (1)

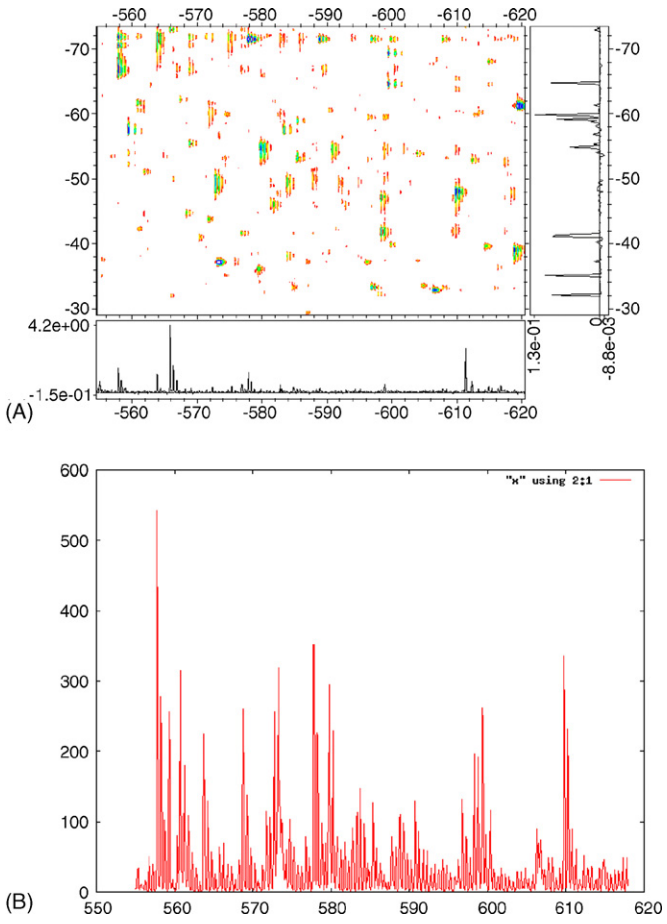

Fig. 1. The visualization of the fragment of a 2D PMF map of human blood plasma peptidome. Panel (A): a color coded 2D map (Sparky visualization tool) *m/z* values—horizontal axis, retention time—vertical axis. Colored spots indicate MS peaks with amplitudes increasing from red to blue. Cross-sections along retention time axis (top) and *m/z* axis (bottom) are shown. Panel (B): a projection of this fragment on the mass-to-charge axis (vertical axis – signal intensity). In this case signals from different peptides can not be resolved. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

  if FOM is less than the value induced by the predefined RL value threshold build the isotopic model for all possible charges and choose one with the smallest FOM
Step 5: calculate the monoisotopic mass of the peptide

3. Algorithm description

The main goal of the algorithm is to reduce the peak list into the list of the monoisotopic peptide masses present in a single LC–MS dataset. In the list of peaks (corresponding to the signals in the spectrum), each peptide is described by several peaks, corresponding to different charge states of the peptide and different isotopic composition. Hence, initially the peak list contains much redundancy. We eliminate this redundancy by determining monoisotopic mass and charge of each signal classified as a peptide.

Major possible application of our approach is the medical diagnosis based on the PFM data. To this aim the contents of series of datasets (lists of peptide masses) have to be compared. This is a highly non-trivial task because of the huge input size.

For this problem we have proposed an efficient heuristic, which allows to process a set of peptides of cardinality up to 200,000.

The main steps of the algorithm are the following.

3.1. Input data and algorithm parameters

An input for the algorithm consists of two datasets: raw and preprocessed MS data. Raw data corresponds to smoothed 2D spectrum in NMRPipe format and preprocessed data corresponds to the list of peaks containing coordinates (mass to charge ratio and retention time), height and volume for each peak. The parameters for peaks are calculated during the peak picking phase.

Both datasets are used further on. At each step peak list is used for the analysis but at some stages the software refers to raw data for the improvement of the analysis of ambiguous cases.

The list of most important parameters of the algorithm include: *max_charge*, i.e., the maximum charge of isotopic cluster which should be considered by the algorithm and the reliability (RL) value threshold. The range of values of *max_charge* parameter is determined by the spectrometer resolution and the range of considered masses. Reliability threshold equals to the required percentage of correctly identified peptide signals. It induces the threshold for the quality of peptide identification, i.e., figure-of-merit (FOM) value (cf. Eq. (1)). In the sequel we classify the group of signals as the isotopic envelope of some peptide only if the FOM of the group is sufficient to guarantee assumed RL value. The FOM thresholds are calculated separately for different peptide charges by manual inspection of algorithm outcomes (data not shown).

3.2. Step 1: noise filtration

Peak picking procedure produces a list of m/z values. At the step of noise filtration the peaks of amplitude below threshold (T) are discarded. The T value is established based on analysis of the distribution of the signal amplitude abundance in a given sample (Fig. 2).

The threshold is fixed to filter out small peaks with comparable height. To this aim we approximate the density function

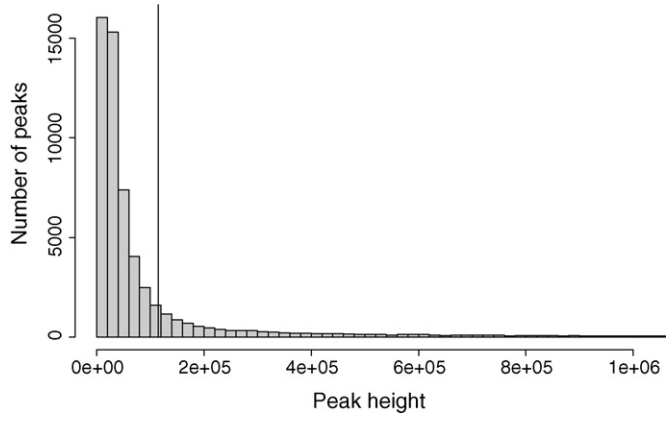

Fig. 2. Distribution of signal abundance in a single blood serum peptidome: the cutoff threshold is calculated to filter out overrepresented small peaks.

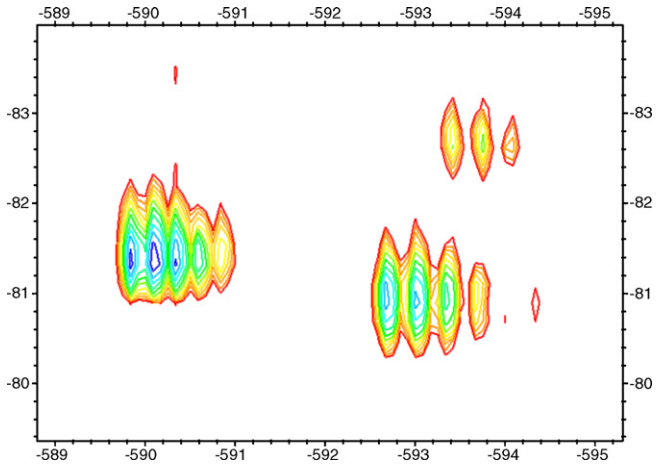

Fig. 3. Enlarged fragment of this figure: examples of isotopic clusters visualized by Sparky. Horizontal axis—m/z, vertical axis—retention time, amplitudes are color coded increasing from red to blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

(f) of peak abundance distribution. The value T corresponds to the signal intensity in which the derivative equals to -1, i.e., $f'(T) = -1$ (Fig. 3).

The procedure can be run globally for the whole analyzed mass-to-charge ratio interval, or locally for small interval to increase the sensitivity of the procedure.

3.3. Step 2: isotopic cluster identification

At this stage our algorithm operates on the list of peaks computed during the preprocessing phase. Peaks are represented by several parameters: m/z value, retention time, height and volume. During peak clustering we take into account general properties of peptide isotopic clusters. We use the *sweeping method* [34] to deal with the 2D data. We scan our 2D input dataset from left to right (direction of increasing mass-to-charge ratios) and examine peaks lying in the vertical stripe of 1 Da width. The position of the right border of the stripe is determined by the mass-to-charge ratio of the actually examined peak (we call it an *active peak*).

We assume that all peaks to the left of the active peak have been already clustered. All peaks with mass-to-charge ratio greater than the active peak will be considered in next steps.

We call a cluster active when its last peak (i.e., the one with the highest mass-to-charge ratio) is in the actually considered stripe. We maintain the data structure containing the list of active isotopic clusters.

Our goal is to assign an active peak to one of active clusters. As the first step we select from the set of active clusters those which could be extended by our active peak.

The criteria for the peak fitting to an isotopic cluster are the following:

- distance between the peak and the cluster (i.e., between the peak and the rightmost peak in the cluster) in the domain of retention time should be smaller than the predefined threshold.

- the shape of the isotopic cluster extended by the active peak should pass user predefined filter (see below). By shape we mean the relative height, positions and the number of peaks in the cluster.

As a filter for the isotopic cluster we investigate here the proportions of the height of two neighboring peaks. We compute the possible extreme values for these proportions by considering polyserine and polyphenyloalanine peptides and we filter out clusters having these proportions outside computed values. In fact these extreme values are further relaxed to encompass sulphur containing peptides. Our algorithm is designed to deal with small peptides. Hence only two possibilities for monoisotopic peak position are considered by the algorithm: either the first or the second peak in the isotopic cluster can be the highest one, and all peaks following it have lower height.

The behavior of the algorithm depends on the number of candidate active isotopic clusters. If there is no candidate cluster to which we can assign our active peak, we form a new cluster containing this peak. When there is only one candidate cluster we extend it with the active peak. In the case when more than one candidate exists we assign the active peak to the cluster whose monoisotopic peak has the highest signal. Such a situation is quite rare but it happens when the signal coming from one peptide is artificially split in the domain of the retention time (cf. Fig. 4).

3.4. Step 3: automated charge determination

We have implemented two versions of this step, simple and fast and a more sophisticated one. The simple version uses only information from the peak spacing in the isotopic clusters as prepared in the previous step and it can be viewed as a variation of the *Z*-score method from [38]. We assume that the charge is simply the reciprocal of the distance between two adjacent peaks in the isotopic cluster. We count results for each possible space interval and choose the most frequent value as a charge.

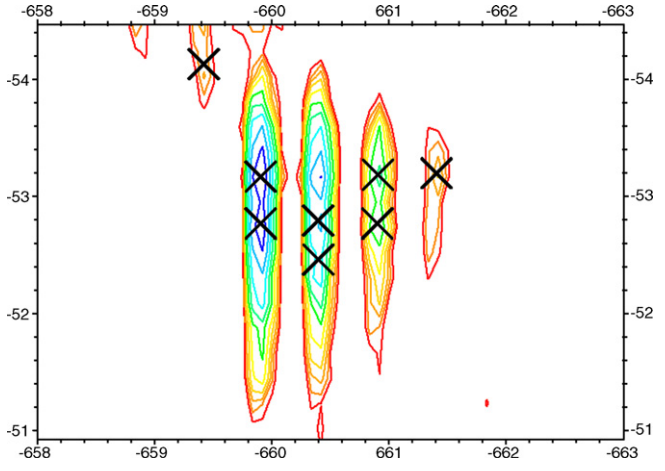

Fig. 4. Artificial peak separation in the retention time domain. Isotopic envelopes visualized by Sparky. Peaks found by NMRPipe are depicted as black 'X'. Horizontal axis—*m/z*, vertical axis—retention time, amplitudes are color coded increasing from red to blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

This method is very fast but also susceptible to errors especially when there are artifacts and split peaks in the spectrum. This method also cannot determine charges for overlapping isotopic envelopes.

Second method is a variation of the method from [29]. It operates on the list of clusters and raw mass spectrum. The original method is designed for 1D spectra. We use here the isotopic clusters found in the previous step to approximate coordinates of isotopic clusters in the spectrum. We perform the projection of the isotopic cluster in the direction of the retention time and use the combination of Patterson and Fourier transform for the *m/z* values of the isotopic cluster to determine the charge. This method can handle the charge up to half of the spectrum resolution.

3.5. Step 4: isotopic model (mass decomposition)

Recall that our procedure is designed to deal with small peptides (up to 5000 Da). For such data the *averagine* model by Senko et al. [30] is not suitable. However to estimate the significance of the cluster we fit its group of peaks to the estimated theoretical isotopic distribution calculated for the given monoisotopic mass. This step is coupled with the mass and charge determination as illustrated in the flowchart (cf. Fig. 5). We start with the assumption that the first visible peak in the isotopic cluster corresponds to the monoisotopic mass. If this

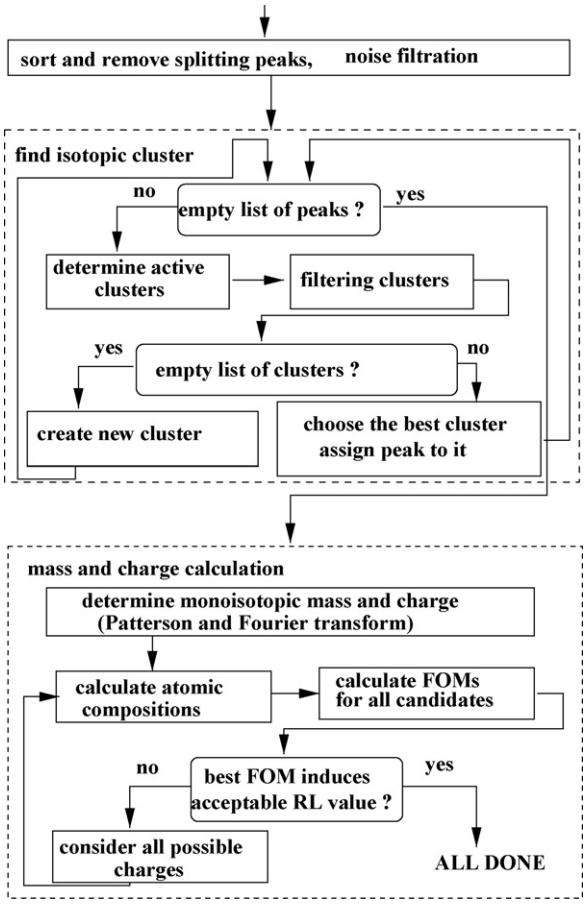

Fig. 5. Flowchart summary of the algorithm.

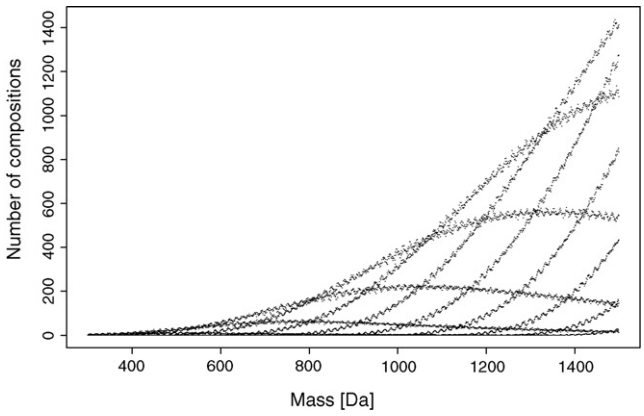

Fig. 6. The distribution of peptides' abundance for a given monoisotopic mass and precision $\epsilon = 0.01$. This graph does not seem to be a function, but it is in fact: for each mass there exists exactly one point corresponding to the number of different atomic compositions. This function behaves very non-continuously: different visible curves arise from interesting combinatorial property of mass decomposition problem, namely, some masses have much larger amount of possible atomic compositions than others.

assumption turns out to be false, the first visible peak in the spectrum is assumed to be the second one (i.e., corresponding to the isotope containing one neutron more).

From the putative mass-to-charge ratio and the charge determined in the previous step we calculate the monoisotopic mass. Then we perform the mass decomposition [28], i.e., we guess the amino acid (and also atomic) composition possible for the given mass. Because the internal standars of known molecular mass is always present in the sample the mass accurancy of the plasma peptide masses is considerable.

Our mass decomposition procedure works as follows. Let m be a monoisotopic mass of a peptide. First, we find candidates for atomic compositions of this peptide: each candidate can be represented as a vector of length 5, storing the numbers of atoms of C, H, N, O and S. Mass of each candidate can differ from m by at most ϵ and has to represent a chain of amino acids.

In order to be able to efficiently find compositions of masses up to M, we perform the following preprocessing: let m_{h} be the mass of the heaviest amino acid considered. We generate all compositions of peptides with mass not exceeding $(M/2) + m_{\mathrm{h}}$, sort them by mass and store in a vector v. To answer a query m, for each element of v we check if there exists an element in v, such that the sum of masses of those two elements differs from m by at most ϵ.

For small peptides our procedure gives a reasonable number of candidate atomic compositions (cf. Table 1 and Fig. 6).

For each candidate atomic composition of a given monoisotopic mass, we calculate theoretical isotopic distribution using dynamic programming technique. Then we fit our experimental data (i.e., isotopic cluster determined in the previous step) to it. To estimate the quality of fit, the widely used figure-of-merit (FOM) value is calculated as follows (cf. [16]):

$$\mathrm{FOM} = \frac{\text{number of points}}{\Sigma[(A_n - ZI_n) + (ZV)^2]} \quad (1)$$

where A_n is the abundance of the nth peak in the theoretical isotopic distribution, I_n the observed signal intensity at the point corresponding to the nth isotopic peak, V the maximum value in the valley between adjacent peaks and Z is the normalization factor. The valley between peaks is the interval from $1/3$ to $2/3$ of the distance between consecutive peaks. The number of points equals the number of values compared (i.e., all peaks and valleys in the isotopic cluster). The normalization factor scales the intensities such that the average intensity from three most abundant peaks in the theoretical distribution equals the average intensity for three corresponding peaks from experimental spectrum. The exception is made for very small masses when the first peak, being the most abundant one, is used to scale the distribution.

The best fit is selected for further analysis. Its FOM has to exceed some user-predefined threshold. This threshold is fixed to guarantee the appropriate reliability (RL) value, i.e., the percentage of correctly identified peptides. The relationship between FOM and RL value is estimated for all possible charges in the way analogous to the THRASH approach [16].

3.6. Step 5: mass and volume calculation

To determine the monoisotopic mass one needs to know the charge of the isotopic cluster and the coordinates of the monoisotopic peak. Both these values are determined in the previous step by the best fit to the theoretical isotopic model (i.e., the fit with the greatest FOM value). The volume, corresponding to the abundance of the peptide, is calculated as the sum of volumes for all peaks in the isotopic cluster. The volume is used later for normalization and classification purposes.

3.7. Postprocessing: aligning masses from different samples

In order to align masses from different samples appropriate clustering algorithm was used. Since the size of the input for the clustering procedure can get quite large (almost 200,000 masses overall in 59 samples considered in the largest dataset), we have decided to implement the following simple and efficient heuristics. The effect of clustering is illustrated in Fig. 7. Notice that besides aligning masses from different samples clustering allows to group masses of the same peptide coming from differently charged ions in one sample.

Input: lists of masses from different samples
Output: list of mass clusters
(1) **while** unclustered masses exist
(2) pick an unclustered mass at random and form a singleton cluster
(3) **while** exists an unclustered mass that fits in a window of predefined size centered at the centroid of that cluster (for small masses ≤ 500 Da predefined window size equals to 0.2 Da and for larger masses this value increases linearly with m/z).

Table 1
The number of candidate atomic compositions for lysozyme peptide mass measured with different precision

Mass	ϵ	Number of candidate compositions
1428.65	0.0001	1
	0.001	14
	0.01	123
	0.1	1157

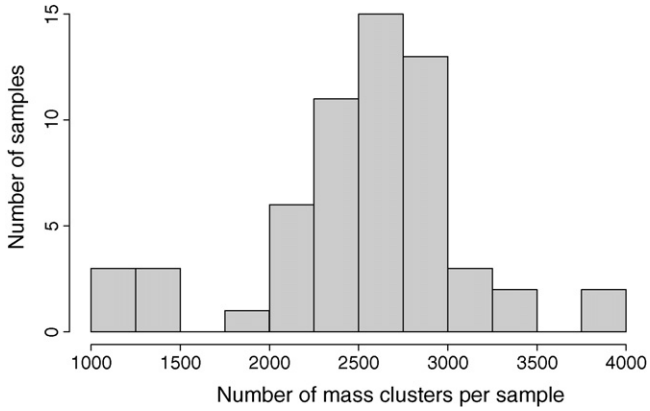

Fig. 7. The distribution of masses per sample (left) and the mass clusters (right). Statistics are calculated on 59 samples.

(4) add to the cluster the mass that is nearest to the centroid along the m/z axis
(5) with probability 1/2 choose equally likely the mass that is farthest from centroid along the m/z axis or along the retention time axis, remove it from the cluster and add to the set of unclustered masses.

3.8. Implementation and availability

Program is written in the C++ language with STL (Standard Template Library). Input/output library for NMRPipe files is written in the C language. We have performed tests of our program under GNU/Linux operating system and on various hardware architectures. It should be easily portable to other operating systems. All program files are downloadable from http://www.sourceforge.net/projects/mz2m. Program works in batch mode without user interaction. There are several command line parameters related to input, output and algorithm operations. Besides the default mode, which calculates the list of masses in a single spectrum, the program can also work in the so called *diff mode*. In this mode program calculates the list of masses which differentiate the two MS spectra (i.e., it calculates the symmetric difference of two sets of masses) (Fig. 8).

4. Results and discussion

4.1. Automated analysis of complex spectra

The goal of the algorithm is to calculate the list of peptides in the sample identified by their mass and retention time. The program has been tested on several datasets. Before starting analysis of complex peptide mixtures many relatively simple samples have been processed for calibration of the whole procedure.

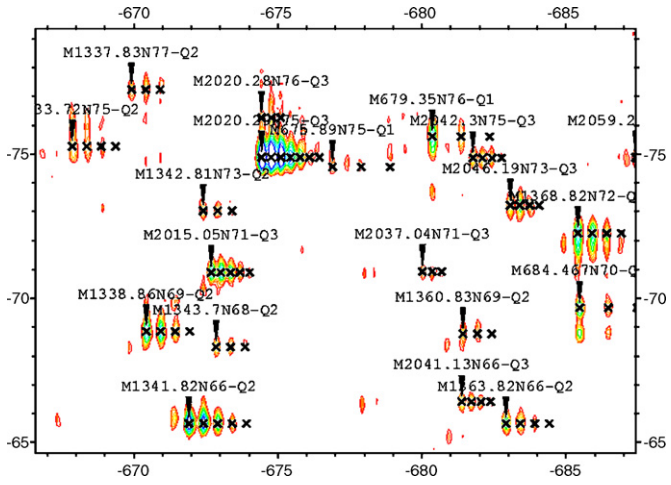

Fig. 8. Masses and charges calculated by our algorithm for the fragment of the spectrum. Peaks are marked as black crosses, small arrow denotes the monoisotopic peak in each isotopic cluster, the monoisotopic mass (M) and charge (Q) are given for each identified peptide.

These samples include tryptic digest of: bovine serum albumin (BSA) (molecular weight 67 KDa), lizozyme (1428.67 Da) and cytochrome *C*.

To demonstrate the application of the LC–MS for large-scale PMF analysis the following sets of highly complex peptide mixtures were processed (cf. Table 2):

(1) blood plasma samples from cystic fibrosis (CF) children and their healthy family members (59 samples),
(2) blood serum samples from colorectal cancer (CC) patients and healthy donors (40 samples).

To estimate the quality of our algorithm several tests have been performed. However, the main goal was to verify the following two aspects which are crucial for medical applications:

- how many peptide signals have been missed by the automated processing?
- how many signals have been interpreted incorrectly?

The best way of validation here was visual insection of the results. For the fragments of PMF of some samples all peptide signals were manually counted and interpreted with the assistance of Sparky visualization tool [13]. The result has been compared to the program output. Table 3 presents the number of interpreted peptide signals and the number of errors. We consider four types of errors: false positives, missing peptides, incorrect charge and incorrect mass calculation. Program misses about 6% of all peptides and returns about 6% peptides with incorrect

Table 2
Masses and peaks statistics

Type of sample	Number of samples	Minimum	Maximum	Mean	Standard deviation
CF (peaks)	59	25.613	108.831	63.032	16.147
CF (masses)	59	1.341	5.244	3.361	821
CC (peaks)	40	57.719	213.178	124.225	53.629
CC (masses)	40	2.657	8.227	5.250	2.250

CF, cystic fibrosis dataset; CC, colorectal cancer dataset.

Table 3
Errors statistics for three manually tested datasets

	Mean	Variance
Manually counted peptide signals	377	14
Correct program outcomes	321	10
False positives	8	7
Incorrect charge	5	3
Incorrect mass	25	1.4
Not found peptides	26	1.4

mass. The incorrect interpretation of the charge is closely related to very strong signal deformation (cf. Fig. 4). We want to emphasize that the peptide list generated by our algorithm contains only about 2% of false positives (possibly experimental and software artifacts).

4.2. Biomarker identification

We have analyzed PMF datasets containing small peptides (up to 3500 Da) obtained using nano-HPLC coupled with LTQ FTICR spectrometer. We sketch here the results obtained for two considered classification tasks. The first one aims at classifying cystic fibrosis patients based on their blood plasma peptidome. The goal of the second task was to identify significant biomarkers for colorectal cancer datasets (i.e., peptides which discriminate well between samples from healthy and diseased people).

Several dimensionality reduction methods have been tested on the obtained datasets [15,36,19]. We present here only the results of Peak Probability Contrast (PPC) method from [35]. Fig. 9 presents an example of a discriminating peptide (i.e., a potential biomarker). In this study we have identified more than 40 statistically significant biomarkers. Mostly, they correspond to peptides having high abundance in the group of healthy patients and in the same time are underrepresented in the group of diseased patients. To estimate the statistical significance of discriminatory peptides we have applied the false discovery rate (FDR) method from [32]. It is worth to mention that in the case when thousands of features in the PMF data set are tested against some null hypothesis, and a number of features are expected to be significant the standard p-value calculation is not a correct approach. In [32] a new method have been proposed for measuring the statistical significance in the case of multiple hypothesis testing. This measure called the q-value is similar to the well known p-value, except it measures the significance in terms of the false discovery rate rather than the false positive rate.

All discovered biomarkers are specified by their monoisotopic mass, charge and retention time and could be further selected for targeted MS/MS identification. This rises the possibility of identifying corresponding proteins and analyzing their role in the investigated pathological process.

For colorectal cancer PFM dataset (40 blood serum samples) we have also identified several statistically significant biomarkers (best four resulted from the PPC method are depicted in Fig. 10). The significance was estimated using the false-discovery-rate approach, i.e., for each peptide we have calculated the probability, that it discriminates well between two random classes.

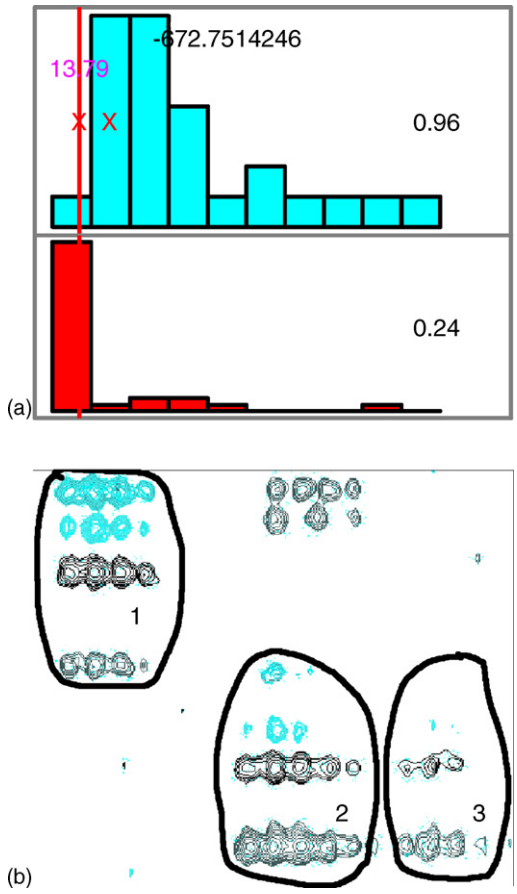

Fig. 9. Results of PPC method on the cystic fibrosis dataset. The left panel shows a histogram of peptide signal amplitudes in the training set for one mass cluster for healthy donors (top) and CF patients (bottom). Cluster centroid m/z value is given (in this case 672.751). The vertical red line corresponds to the calculated discriminating threshold: the blood serum of 96% of healthy donors contains the investigated peptide amplitude above this threshold but only 24% of CF patients blood plasma shows larger amplitude of this peptide. Right panel: appropriate fragments of four PMF datasets are visualized (two healthy colored blue, two diseased colored black), three different peptide signals are visible (they differ by m/z and the retention time), peptide 3 discriminates well between groups. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

This probability should be small enough to call a peptide significant. The interesting phenomena can be observed in Fig. 10, namely the FDR value as a function of the PPC threshold is not monotonic: there exist several peptides having high PPC threshold but also high FDR value. According to us these peptides are spurious biomarkers and should not be selected for further analysis.

4.3. Time and space complexity

We are aware of the other approach to analyze 2D LC–MS spectrum [3]. The authors have decided to process the data in the distributed environment. The spectrum is analyzed there in parallel, for each retention time separately. This seemed to be necessary because of the high complexity of the problem. In our method we process a 2D spectrum sequentially and the time of the whole procedure is comparable to the time of processing ca.

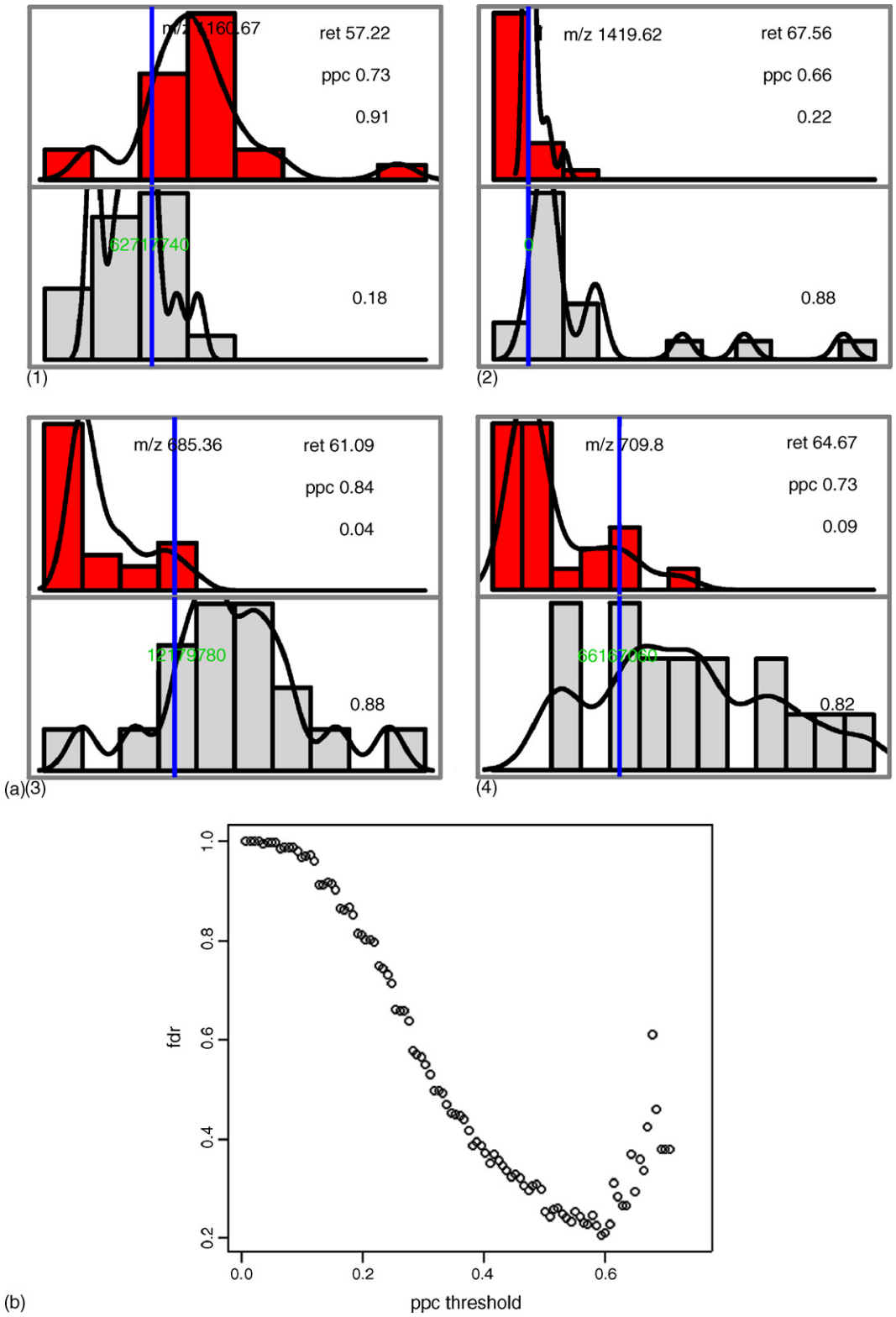

Fig. 10. Results of PPC method on the colorectal cancer example. Each of the four panels in the right figure shows a histogram of peptide intensities in the training set for one mass cluster for healthy donors (top) and cancer patients (bottom). Cluster centroid coordinates: *m/z* and retention time value is given. Additionally we estimated the density of signal abundance and calculated the PPC threshold as before. Right panel shows the false-discovery-rate (FDR) value as a function of the PPC threshold.

two lines (retention times) in the previous approach. Our algorithm operates on the sorted peaks list (of length n) and calculates the list of masses (i.e., peptides) of length k. Assuming that every vertical strip intersects constant number of isotopic clusters (in practice this number is less than 10), we can estimate time complexity of the algorithm to be $O(n(\log n + \log k))$. Memory requirements can be bounded by $O(n + k)$. Effectivity tests were performed to find the execution time and memory requirements.

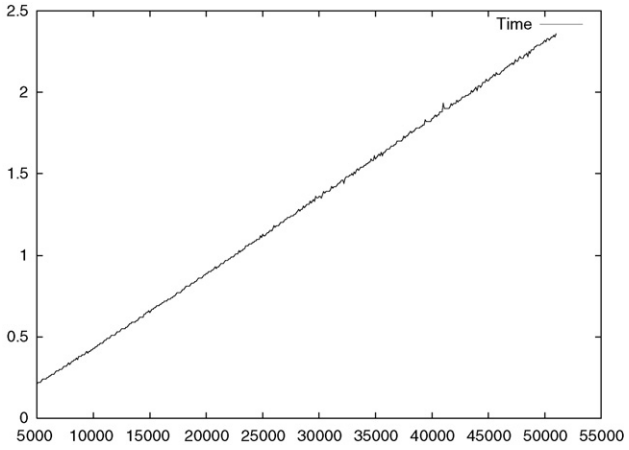

Fig. 11. The processing time of the algorithm as a function of input size. The time needed to process a sample scales linearly with the sample size. Vertical axis corresponds to the time (in seconds) and the horizontal axis corresponds to the input size (i.e., number of peaks).

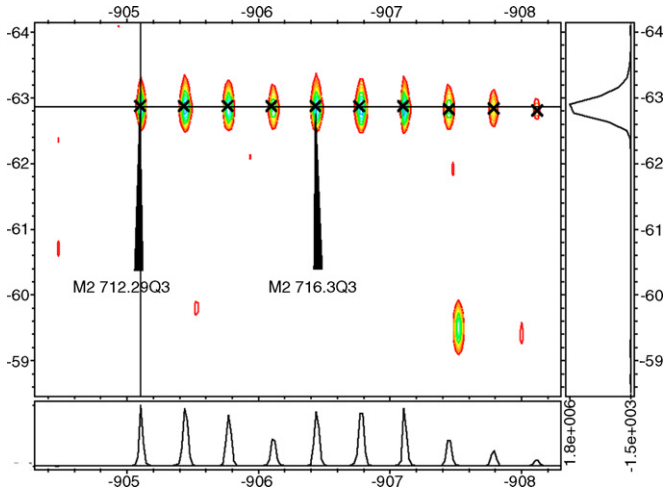

Fig. 12. Application in differential proteomics: our program correctly identified two isotopic clusters resulting from the mixture of $^{16}O/^{18}O$ labeled peptides. Horizontal axis—m/z, vertical axis—retention time, projections onto both axes are presented at the bottom and right side of the figure.

Experiments were made on double Intel Xeon 2.8 GHz PC with 2GB RAM memory, Linux version 2.6.8 (cf. Fig. 11).

5. Conclusions

In this paper we have described a new algorithm for automated peptide identification which is designed to deal with 2D data resulting from LC–MS. The algorithm is intended to analyze already preprocessed MS spectra: it requires prior noise-reduction, signal-smoothing, normalization and peak picking. There exist tools specialized for this preliminary analysis, e.g. [4], but we strongly suggest to use the multidimensional spectral processing system NMRpipe [11] which provides automated peak detection in 1D–4D, including options for identifying peaks due to random noise and truncation artifacts. Our algorithm can be viewed as a substantial generalization of the THRASH method [16] for two dimensions. It can efficiently process LC–MS spectra with sensitivity sufficient for medical applications (e.g., biomarkers search). The ability of correct interpretation of peptide signals is crucial in classification for medical diagnostics (screening and prognostic tests).

An interesting open question is how to deal with more dimensions. A challenging problem is to design the framework for multidimensional mass spectrometry based on different separation techniques [18,33]. It would be interesting to verify the applicability of the recent progress in the field of high dimensional computational geometry [14] to this task.

Other applications of our algorithm include protein identification [5] and differential proteomics. The algorithm can be used to calculate masses for peptides coming from a given digested protein and automatically detect assumed modifications like post-translational modifications, crosslinked species, etc. The method can also be useful for LC–MS based differential proteomics, in which quantitative comparison of protein levels in two samples is made possible by labeling peptides in one of the samples by stable isotope [2,17]. The example of application of our program to the peptide mixture labeled either with ^{16}O or ^{18}O. The program clearly allows to differentiate ^{16}O and ^{18}O labeled species in an automated way and to quantitate peptide ratio. We plan to extend the functionality of our tool to perform this kind of analysis too (cf. Fig 12).

Acknowledgments

This work was supported by Polish Ministry of Education and Science grants KBN-8 T11F 021 28 and PBZ-KBN-088/P04/2003.

References

[1] B.-L. Adam, Y. Qu, J.W. Davis, M.D. Ward, M.A. Clements, L.H. Cazares, O.J. Semmes, P.F. Schellhammer, Z. Yasui, Y. Feng, G.L. Wright Jr., Cancer Res. 62 (13) (2002) 3609.
[2] S.N. Twigger, B.D. Halligan, R.Y. Slyper, J. Am. Soc. Mass Spectrom. 16 (2005) 302.
[3] M. Biskup, Master Thesis, UW-Vrije, 2004.
[4] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, Anal. Chem. 78 (3) (2006) 779.
[5] L. Li, D. Craft, A. Doucette, J. Proteome Res. 1 (2002) 537.
[6] E.P. Diamandis, Clin. Chem. 49 (2003) 1272.
[7] E.P. Diamandis, Mol. Cell. Proteomics 3 (2004) 367.
[8] S. Ryu, T.G. Hamilton, E. Foss, Y. Mao, D. Radulovic, S. Jelveh, A. Emili, Mol. Cell. Proteomics 3 (2004) 984.
[9] L.A. Liotta, E.F. Petricoin, Trends Biotechnol. 20 (12) (2002) s30.
[10] E.P. Diamandis, J. Natl. Cancer Inst. (5) (2004) 353.
[11] G.W. Vuister, G. Zhu, J. Pfeifer, F. Delaglio, S. Grzesiek, A. Bax, J. Biomol. NMR 6 (1995) 277.
[12] V.A. Fusaro, J.H. Stone, Clin. Exp. Rheumatol. (6 Suppl. 32) (2003) S3.
[13] T.D. Goddard, D.G. Kneller, SPARKY 3, TR University of California, San Francisco.
[14] J.E. Goodman, J. O'Rourke, Handbook of Discrete and Computational Geometry, CRC Press LLC, Boca Raton, 2004.
[15] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning, Springer, 2001.
[16] D.M. Horn, R.A. Zubarev, F.W. McLafferty, J. Am. Soc. Mass Spectrom. 11 (4) (2000) 320.
[17] D. Figeys, I.L. Steward, T. Thomson, Rapid Commun. Mass Spectrom. 15 (2001) 2456.
[18] G.M. Janini, T.P. Conrads, T.D. Veenstra, H.J. Issaq, K.C. Chan, J. Chromatogr. B 817 (1)(5) (2005) 35.

[19] J.C. Langford, J.B. Tenenbaum, V. de Silva, A Global Geometric Framework for Nonlinear Dimensionality Reduction, 2000.
[20] E.C. Kohn, L.A. Liotta, E.F. Petricoin, JAMA 286 (2001) 2211.
[21] J. Li, Z. Zhang, J. Rosenzweig, Y.Y. Wang, D.W. Chan, Clin. Chem. 48 (8) (2002) 1296.
[22] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, Lancet 359 (9306) (2002) 572.
[23] E.F. Petricoin, K.C. Zoon, E.C. Kohn, J.C. Barrett, L.A. Liotta, Drug Discov. 1 (9) (2002) 683.
[24] E.F. Petricoin, L.A. Liotta, Clin. Chem. 49 (2003) 533.
[25] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, Lancet 359 (9306) (2002) 572.
[26] Y. Qu, B.-L. Adam, Y. Yasui, M.D. Ward, L.H. Cazares, P.F. Schellhammer, Z. Feng, O.J. Semmes, G.L. Wright Jr., Clin. Chem. 48 (10) (2002) 1835.
[27] A.J. Rai, Z. Zhang, J. Rosenzweig, Ie.-M. Shih, T. Pham, E.T. Fung, L.J. Sokoll, D.W. Chan, Arch. Pathol. Lab. Med. 126 (12) (2002) 1518.
[28] J.A. Kowalak, S.C. Pomerantz, J.A. McCloskey, J. Am. Soc. Mass Spectrom. 4 (1993) 204.
[29] M.W. Senko, S.C. Beu, F.W. McLafferty, J. Am. Soc. Mass Spectrom. 6 (1) (1995) 52.
[30] M.W. Senko, S.C. Beu, F.W. McLafferty, J. Am. Soc. Mass Spectrom. 6 (4) (1995) 229.
[31] P.R. Srinivas, M. Verma, Y. Zhao, S. Srivastava, Clin. Chem. 48 (8) (2002) 1160.
[32] J.D. Storey, R. Tibshirani, PNAS 100 (16) (2003) 9440.
[33] A.A. Shvartsburg, E.F. Strittmatter, R.D. Smith, K. Tang, F. Li, Anal. Chem. 77 (9) (2005) 6381.
[34] C.E. Leiserson, T.H. Cormen, R.L. Rivest, Introduction to Algorithms, The MIT Press/McGraw-Hill, 1990.
[35] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, Q.-T. Le, Bioinformatics 20 (17) (2004) 3034.
[36] J.B. Tenenbaum, V. de Silva, Adv. Neural Inf. Process. Syst. 15 (2003)
[37] C.J. Kemp, H. Zhang, X. Li, E.C. Yi, R. Aebersold, Mol. Cell. Proteomics 4 (2005) 1328.
[38] Z. Zhang, A.G. Marshall, J. Am. Soc. Mass Spectrom. 9 (3) (1998) 225.
[39] W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm, J.S. Kovach, Proc. Natl. Acad. Sci. U.S.A. 100 (25) (2003) 14666.