EuPA OPEN PROTEOMICS 4 (2014) 87–100

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://www.elsevier.com/locate/euprot

Towards automated discrimination of lipids versus peptides from full scan mass spectra

Piotr Dittwald[a,b], *Trung Nghia Vu*[c,d], *Glenn A. Harris*[e], *Richard M. Caprioli*[e], *Raf Van de Plas*[e,1], *Kris Laukens*[c,d,1], *Anna Gambin*[b,f,1], *Dirk Valkenborg*[g,h,i,*,1]

[a] *College of Inter-faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Warsaw, Poland*
[b] *Institute of Informatics, University of Warsaw, Warsaw, Poland*
[c] *Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium*
[d] *Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp/Antwerp University Hospital, Edegem, Belgium*
[e] *Mass Spectrometry Research Center and Departments of Biochemistry, Chemistry, Pharmacology, and Medicine, Vanderbilt University, Nashville, USA*
[f] *Mossakowski Medical Research Centre, Polish Academy of Sciences, Warsaw, Poland*
[g] *Applied Bio & Molecular Systems, VITO, Mol, Belgium*
[h] *Center for Proteomics, Antwerp, Belgium*
[i] *Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium*

ARTICLE INFO

Article history:
Available online 27 May 2014

Keywords:
Lipidomics
Peptidomics
Bioinformatics
Machine learning
Lipid/peptide classification
Lipid centrifuge

ABSTRACT

Although physicochemical fractionation techniques play a crucial role in the analysis of complex mixtures, they are not necessarily the best solution to separate specific molecular classes, such as lipids and peptides. Any physical fractionation step such as, for example, those based on liquid chromatography, will introduce its own variation and noise. In this paper we investigate to what extent the high sensitivity and resolution of contemporary mass spectrometers offers viable opportunities for computational separation of signals in full scan spectra. We introduce an automatic method that can discriminate peptide from lipid peaks in full scan mass spectra, based on their isotopic properties. We systematically evaluate which features maximally contribute to a peptide versus lipid classification. The selected features are subsequently used to build a random forest classifier that enables almost perfect separation between lipid and peptide signals without requiring ion fragmentation and classical tandem MS-based identification approaches. The classifier is trained on *in silico* data, but is also capable of discriminating signals in real world experiments. We evaluate the influence of typical data inaccuracies of common classes of mass spectrometry instruments on the optimal set of discriminant features. Finally, the method is successfully extended towards the classification of individual lipid classes from full scan mass spectral features, based on input data defined by the Lipid Maps Consortium.

* *Corresponding author at*: Applied Bio & Molecular Systems, VITO, Mol, Belgium. Tel.: +32 32653387.
E-mail addresses: piotr.dittwald@mimuw.edu.pl (P. Dittwald), dirk.valkenborg@vito.be (D. Valkenborg).
[1] Authors share senior authorship.
http://dx.doi.org/10.1016/j.euprot.2014.05.002

1. Introduction

In analytical chemistry, and specifically in mass spectrometry, instrumental developments continuously push the boundaries of sensitivity and resolution. This boost in the level of spectral detail makes it increasingly feasible to learn aspects of the identity of a compound directly from the spectrum, which is particularly valuable when complex mixtures are analyzed. Although fractionation techniques such as liquid chromatography are widely used to reduce the complexity of mass spectral data, they rarely attain a perfect separation in which molecular classes of interest are isolated from other molecular components in the sample (e.g. separation of peptides from lipids in a peptidomics study) [1]. Additionally, in certain studies the use of hyphenated techniques is incompatible or impractical (e.g. mass spectral imaging [2–4] of bioactive peptides).

In this work we use some of the additional information provided by contemporary mass spectrometers in terms of mass resolution to provide a computational answer to the separation challenge. Specifically, we have developed an automated method to discriminate between peptide and lipid peaks observed in full scan mass spectra. After investigating, in a generic sense, the isotopic behavior and corresponding masses of different molecular classes, we propose a computational approach that offers a preliminary interpretation of the molecular content of a full scan mass spectrum, without the need for ion fragmentation and classical tandem MS-based identification. This manuscript demonstrates discrimination between polypeptide peaks and lipid peaks in a mass range where both classes co-occur, by using features extracted from the isotope distribution and masses associated with each observed isotope variant. It exceeds the performance of typical rules-of-thumb, such as examining the mass defect of ions, and it does so in an automated way. Although such rules-of-thumb are common in the mass spectrometry community, they typically have not been thoroughly investigated by means of a high-throughput computational analysis. Our aim is to present a rigorous validation of such rules in an *in silico* analysis and to extend them with more powerful heuristics where possible.

The presented work is similar in spirit to the methods of Kirchner et al. [5] and Bruce et al. [6] to discern the degree of phosphorylation of a peptide. Both papers exploit a predefined mass defect caused by the phosphate group. In this paper however, we propose a generic approach that searches for the optimal set of isotope features extracted from representative peptide and lipid databases to enable the discrimination between peptide and lipid classes. The approach also delivers a classification model on the basis of those features. More specifically, we employ a random forest classifier [7,8], a fast and effective multi-classification tool that is based on decisions made by a large set of randomly generated classification and regression trees (CARTs). This approach allows us to investigate the importance of different mass spectrometric features as input variables for the peptide-vs.-lipid classifier. Additionally, we bring these theoretical findings in relation to empirical mass spectrometry measurements by modeling how the expected data inaccuracy of real instruments affects the optimal features for interpretation.

The proposed methodology is further evaluated by an extensive simulation study and a controlled MS experiment on peptides and lipids. Finally, we introduce an extension of the method, which is capable of assigning to the input data, the lipid class probabilities in line with the classes defined by the Lipid Maps Consortium [9]. With regards to isotope-derived information, the BRAIN method proposed by Claesen et al. [10] is key to enabling a theoretical study of isotopic features to discriminate molecular classes. Although many algorithms are available to calculate the isotope distribution of a molecule [11–14], BRAIN provides center-masses (i.e. average masses of isotope variants with the same number of neutrons) in addition to the isotopic peak intensities (see also [15,16]). We use the BRAIN algorithm to produce the theoretical (aggregated) isotopic distribution for a chemical formula. As we will show, access to exact masses of the isotopic variants is an important component for the classification between lipids and peptides when only full scan information is available.

A method that is able to recognize lipids from peptides in full scan mass spectra is particularly useful in specific applications. For example, it can be applied to deconvolve bio-molecular images obtained via MALDI-based imaging mass spectrometry experiments. Such a spatial measurement could potentially be automatically separated into a set of ion images presumed to be lipids, and a set of probable peptide ion images. An automated interpretation would significantly reduce the complexity of analyzing the massive information content accumulated in these types of experiments and it would also enhance the ability of researchers to interpret the biology, which is sometimes obscured by the sheer amount of data collected. Another application would be to start driving the process of annotating a compound at the full scan level, which is nowadays usually dependent on the availability of fragmentation spectra. The probable molecular class annotation delivered by our method can be used to provide a more directed post-acquisition workflow. For example, a presumed peptide ion could be automatically passed on to a database search strategy, whilst a probable lipid could be sent directly to ChemSpider [17], Lipid Maps [9], Metlin [18] or MassBank [19] for annotation. A third application could be a classification workflow embedded in the instrument to help guide data-dependent MS/MS experiments, enabling real-time and on-the-fly selection of the optimal fragmentation strategy per molecule class while the full (parent) mass spectrum is still being collected.

2. Materials and methods

This section introduces the lipid and peptide databases that were used to generate the virtual mass spectra to train the classification model. It also describes the real mass spectral measurements used to test the classification performance on a wet-lab lipid–peptide mix. This is followed by a detailed description of the classification methods employed and the details pertaining to building the model.

2.1. *In silico data (training set)*

Two biological databases were used to explore discriminatory features between peptides and lipids in general. Lipid information was obtained from the Lipid Maps gateway database [9] (downloaded September 2011). To avoid a bias in the classification, all eight lipid classes defined by the Lipid Maps Consortium were part of the lipid-representative set. These classes are fatty acyls (#913; FA), glycerolipids (#400; GL), glycerophospholipids (#1415; GP), sphingolipids (#1167; SP), sterol lipids (#604; ST), prenol lipids (#442; PR), saccharolipids (#76; SL), and polyketides (#1296; PK), with the number of species in each class and their abbreviated class label indicated in parentheses. Thus, all eight lipid classes were taken together to represent one general class of lipids. In the later extension, lipid classes are considered separate so that lipid subclass-specific annotation becomes possible. The classifiers are conceived to discriminate lipids and peptides with a mono-isotopic mass below 2.8 kDa. This constraint on the molecular mass is necessary to ensure that the classification focuses on the mass range where peptide–lipid discrimination is relevant, avoiding obvious classification rules such as, e.g. all molecules above 2.8 kDa are peptides. In total 6313 lipids were included in the study. Only lipids and peptides with a molecular mass below 2.8 kDa were considered, which in the case of lipid molecules amounts to 99.1% of the entries found in Lipid Maps.

The peptide information was retrieved from the Human Uniprot protein database [20]. The extracted protein sequences were tryptically digested *in silico*, using the *digest* function from the OrgMassSpecR package available in the CRAN repository (allowing for no missed cleavages). The resulting database contained 263,897 tryptic peptides with a mono-isotopic mass below 2.8 kDa. To keep the computation efficient and to avoid biased classifiers due to unbalanced training sets, only a random sub-sample of 6313 peptides is used for training so that the number of entries is comparable to the lipid database. To ensure that this selection procedure does not cause bias, we have repeated the classification training five times with five different peptide selections. The performance statistics over the five peptide-representative databases are shown in Section 3.

For each lipid or peptide entry in the collected training set, we calculated the aggregated isotope distribution and the corresponding center-masses using the BRAIN method [10]. For this purpose, the BRAIN software package [21] was extended with additional chemical elements that occur in lipids, namely fluorine, bromine, phosphorus, chlorine, sodium, and iodine. The theoretical isotope distribution was restricted to the first three consecutive aggregated isotope peaks, because for light molecules such as those used in this application, these isotope variants are typically the most prominent. It should be noted that in the theoretical data set only the protonized variants of the molecular species were considered whilst disregarding possible modifications and other adduct formations.

2.2. *Real MS data (test set)*

The *in silico* data in the training set, consisting of thousands of curated peptide and lipid examples, can be used to build a theoretical lipid-vs.-peptide classifier. By itself such a classifier has value by revealing, on a theoretical level, the features that allow discrimination between lipid and peptide peaks. However, in this study we also want to gauge the performance of such a classifier on mass spectral measurements. The goal is to assess how well the features and classifier that are built on clean curated data holds up in the presence of real measurement conditions and the various noise sources that accompany them. To this end, we created a wet-lab mixture of known peptide and lipid species, which is then measured via MALDI-TOF MS to produce a mass spectrum. Since we know the mixture and the species involved, we have a gold standard for the identity of several peaks in the spectrum. Therefore, if we apply the classifier to this mass spectral measurement, it can be used as a test set to verify classification performance.

As indicated in Table 1, the mixture consists of five peptide species and seven lipid species. The peptides include Kemptide (mono-isotopic mass: 771.472 Da), PKC substrate (828.541 Da), ACTH 4-11 (1089.518 Da), Glu1 Fibrinopeptide B (1569.67 Da), and ACTH 18-39 (2464.191 Da). The lipids contain seven distinct species, but in order to mimic experimental conditions further, some were chosen to be isobaric. This essentially reduces the number of distinct lipid masses in the mixture to three. The lipid mono-isotopic masses are 759.578 Da for PC 18:1(9z)/16:0 and PC 16:0/18:1(9z), 785.593 Da for PC 18:1(9z)/18:1(9z), PC 18:1 (9trans), and PC 18:1(6z)/18:1(6z), and finally 761.593 Da for PC 16:0/18:0 and PC 18:0/16:0. The lipid–peptide mixture was hand-spotted on a target plate and coated with sinapinic acid as the matrix. The spot was then measured in a Waters Synapt G2 mass spectrometer (Waters Corporation, Milford, MA) that has been fitted with a MALDI source. The instrument was run under standard supplied manufacturer settings in "Resolution Mode" of the time-of-flight mass analyzer. The resulting spectrum ranges from *m/z* 400 to 3000 spanning 158,701 bins, which amounts to an average bin size of *m/z* 0.016. Due to the relatively high mass resolution of this instrument, the ion peaks in the spectrum are generally isotopically resolved across its mass range. Besides the protonated versions of the species mentioned above and their isotopic variants, there are several other peaks present in the spectrum as can be seen in Supplementary Figure S1. These additional peaks for which we have no gold standard annotation to compare against, are for example reporting adducts (other than M+H) of the analyte species or matrix molecule species.

The extraction of the isotope distributions from the spectrum starts with the detection of mono-isotopic peaks and their charge states using the YADA software [22]. All parameters in YADA are set to default except for the minimum intensity of the mono-isotopic peaks, which was set to 5000. The list of mono-isotopic peaks of the spectrum was used as input for the detection of other peaks in their isotope distributions. In order to reduce spectral complexity, only local maxima were retained to characterize peaks. Local maxima below an intensity of 400 were discarded. The local maxima were centroided by using the mid-point of the peak envelope instead of the *m/z*-value of the apex of the peak. The mono-isotopic peak list was used as a target list to find the consecutive isotope peaks that were separated by 0.95–1.05 Da. Since data prepossessing and peak extraction is not part of

Table 1 – Five peptide species and seven lipid species contained within the test set mixture.

Name	Concentration in pmol/μL	Formula	Avg. MW	Mono mass	Mono mass with H^+	Mono mass with Na^+
Peptides						
Kemptide	2.15	$C_{32}H_{61}N_{13}O_9$	771.918	771.472	772.479	794.461
PKC substrate	2.1	$C_{34}H_{68}N_{16}O_8$	829.016	828.541	829.548	851.53
ACTH 4-11	1.55	$C_{50}H_{71}N_{15}O_{11}S$	1090.268	1089.518	1090.526	1112.508
Glu1 Fibrinopeptide B	1.05	$C_{66}H_{95}N_{19}O_{26}$	1570.592	1569.67	1570.677	1592.659
ACTH 18-39	0.7	$C_{112}H_{165}N_{27}O_{36}$	2465.701	2464.191	2465.199	2487.181
Lipids						
PC 18:1(9z)/16:0	0.95	$C_{42}H_{82}NO_8P$	760.076	759.578	760.586	782.568
PC 18:1(9z)/18:1(9z)	0.95	$C_{44}H_{84}NO_8P$	786.113	785.593	786.601	808.583
PC 16:0/18:0	0.95	$C_{42}H_{84}NO_8P$	762.092	761.593	762.601	784.583
PC 18:1 (9trans)	0.95	$C_{44}H_{84}NO_8P$	786.113	785.593	786.601	808.583
PC 18:0/16:0	0.95	$C_{42}H_{84}NO_8P$	762.092	761.593	762.601	784.583
PC 16:0/18:1(9z)	0.95	$C_{42}H_{82}NO_8P$	760.076	759.578	760.586	782.568
PC 18:1(6z)/18:1(6z)	0.95	$C_{44}H_{84}NO_8P$	786.113	785.593	786.601	808.583

the proposed concept, other peak picking algorithms could be employed instead.

2.3. *Algorithmic approach*

The general concept is illustrated in Fig. 1. The optimal classification model to theoretically discriminate lipids from peptides is selected based on an *in silico* study (left panel). This study evaluates typical errors on mass and spectral accuracy that correspond to different mass spectrometry types and assesses the impact of these errors on the feature sets that drive the classification. The workflow on the right-hand side of Fig. 1 illustrates how a classifier can be used as an automated procedure to discern lipids from peptides in measurements that were not part of the training set. Note that the feature selection and classifier training is only based on *in silico* generated mass spectra, ensuring that the training and testing phase of the classifier are completely independent from each other.

2.3.1. *Feature set selection*

To design a robust and efficient classifier, we create a list of features based on information derived from theoretical isotope distributions and their exact center-masses. More precisely, not only the isotope masses and intensities are used as features, but also the fractional part of these masses, and mass differences between consecutive isotope peaks, etc. The considered features are as follows:

- *mass.1*: exact mass of first isotope peak;
- *mass.2*: exact mass of second isotope peak;
- *mass.3*: exact mass of third isotope peak;
- *mass.frac.1*: fractional part of mass of first isotope peak;
- *mass.frac.2*: fractional part of mass of second isotope peak;
- *mass.frac.3*: fractional part of mass of third isotope peak;
- *mass.diff.21*: difference between second and first isotope masses;
- *mass.diff.32*: difference between third and second isotope masses;
- *iso.ratio.21*: ratio of intensities of second and first isotope peaks;
- *iso.ratio.31*: ratio of intensities of third and first isotope peaks.

For the last two items, the intensities of the isotope distribution were normalized to the intensity of the mono-isotopic peak. This normalization accounts for systematic multiplicative noise and avoids scaling when experimental data with absolute peak intensities is provided.

As we stated earlier, the objective of this research is a classification model that can discern lipids and peptides in mass spectrometry experiments. Since classification rules learned from theoretical data are not necessarily transferable to a real experimental example, direct use of the theoretical information without taking into account mass uncertainty and peak intensity noise is ill-advised. The performance and optimal feature set of the classifier will change as a function of resolving power and spectral accuracy of a particular instrument. Therefore, we conduct a sensitivity analysis to reveal which features are best to discriminate between lipids and peptides in a realistic setting. To accomplish this, the theoretical peaks provided by the BRAIN method are used to produce virtual spectra that resemble spectra from commercially available mass spectrometry instruments. For example, the center-masses of the theoretical isotope distribution are rounded to the nearest value of 5, 4, 3, 2, and 1 decimal digits. Considering the mass range of lipids, these induced errors roughly correspond to the mass accuracy of commercial mass spectrometers consistent with FTICR, Orbitrap, TOF, ion trap and quadrupole analyzers, respectively. In order to explore the sensitivity of the classifier with respect to the error on the isotopic peak intensities, normally distributed noise with mean one and different standard deviations σ are multiplied by the occurrence probabilities of the theoretical isotope distribution. The standard deviation σ takes values of 0.01, 0.1, 0.2, and 0.3, reflecting commonly observed intensity errors. Since the probability of drawing non-positive values from those distributions is small (e.g. approximately 0.00043 for $\sigma = 0.3$), negative values need not be explicitly removed to give a good approximation. In this simulation scheme a homoscedastic error structure is assumed that perturbs all the peak intensities of the theoretical isotope distribution with an

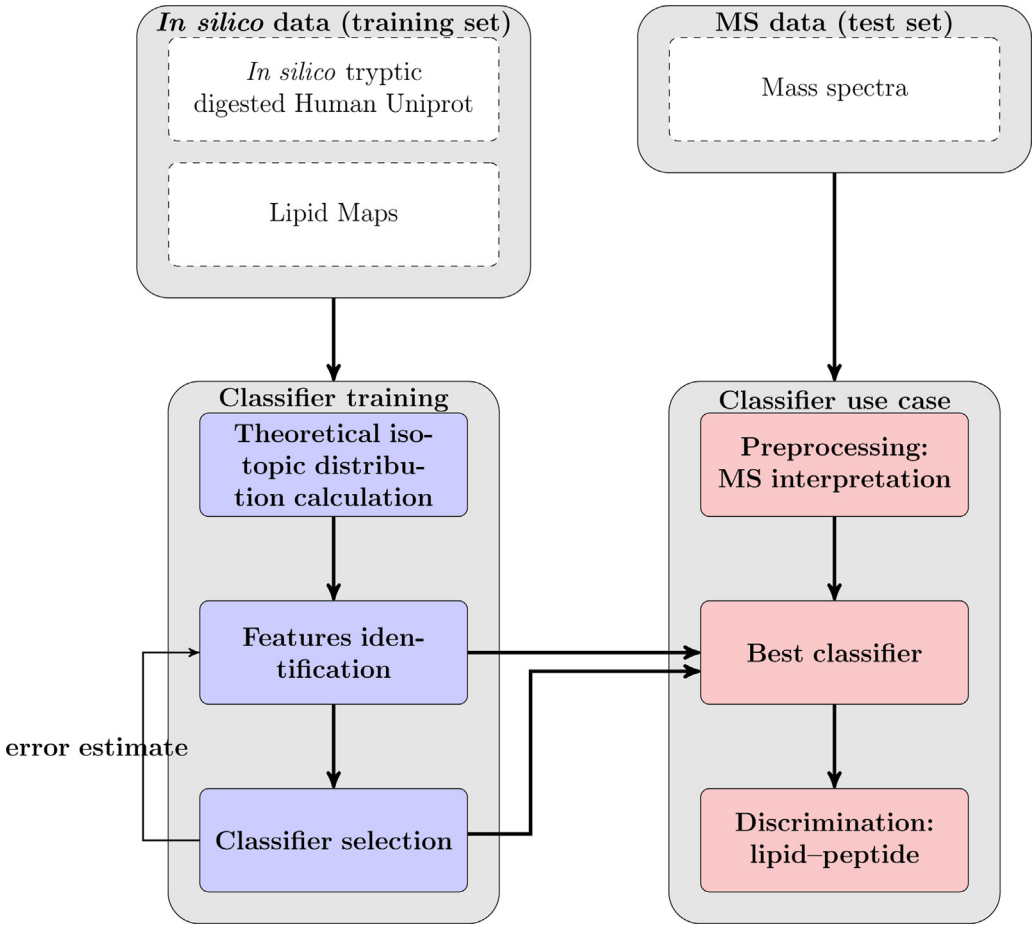

Fig. 1 – Overview of the discrimination algorithm: classifier training scheme (left) and classifier use case (right).

error of the same variance. In total 6×5 different virtual spectra data sets are generated from the lipid and peptide training set. Each of these 30 data sets has its own combination of mass and intensity noise to find an optimal feature set and classifier for. For each generated spectra set, we assess the importance of a feature in the classification decision by examining its Gini index and misclassification rate. These performance metrics will be further explained later on, but a detailed overview is also available in [8].

2.3.2. *Random forest classifier*

For the purpose of discrimination between different molecular classes, we make use of a random forest (RF) classifier [7]. This classifier type uses the idea of aggregating the responses of many classifiers (built from perturbed versions of the training set) into a single classification answer. Specifically, the RF classifier constructs an ensemble of classification trees and makes a final decision based on a majority vote. Each of the constructed trees returns a single classification decision. The final RF decision is then chosen as the most popular among these single decisions. The majority of the votes also shows the strength of the final classification and can be seen as a probability of belonging to a certain class, i.e. the number of trees pointing to this class divided by the total number of built trees. The RF classifier is constructed in such a way that a value close to one indicates that the observed molecule is of peptide nature, while a value close to zero means that it is of lipid nature. Values between zero and one thus represent cases with varying degrees of uncertainty about the molecular class. We use out-of-bag (OOB) error estimates as an error measure [23]. In addition, error rates obtained through cross-validation are presented in Supplementary Table S1. It should be noted that the RF classifier entails a stochastic component. As a result, a training phase repeated on the same data set will produce small fluctuations in the misclassification rates. For this reason it is sensible to investigate the global trend of the misclassification rate, rather than to scrutinize some arbitrary rates.

Whereas the previous paragraph describes the RF classifier to discern peptides from a global lipid class, a second RF classifier can be trained to perform sub-categorization once a decision has been made about the molecular species. The RF training phase for this multi-class classifier uses the same virtual spectra sets from the *in silico* lipid database, but presents them as separate classes rather than as a single lipid class. The mass and intensity error models are analogous to what was used in the two-class peptide-vs.-lipid classifier.

After the training phase of the lipid-vs.-peptide and multi-class classifiers on the virtual spectra sets, an optimal model based on the selected feature set is available for each of the mass and intensity error combinations.

3. Results

The results are divided into four parts. In the first part, we appraise the mass defect rule-of-thumb [24], which is equivalent to looking only at the fractional part of an ion mass to determine the molecular class of an ion peak (cfr. Kendrick maps). The second part provides a theoretical basis for this single-feature rule-of-thumb by means of a sensitivity analysis on the 6×5 virtual spectrum sets. Besides the fractional part of the mass, we also investigate the importance of other isotopic features in terms of the instrument capabilities. For this purpose, we evaluate the an optimal classifier and corresponding feature set for each level of mass resolution and spectral accuracy. The third part elaborates on the training of the multi-class classifier to determine subclasses within the lipid category and it visually illustrates the distribution of isotopic characteristics for each of the lipid classes defined by the Lipid Maps Consortium. Section 3.4 applies the trained classifiers on a real mass spectrum and assesses the performance against a gold standard. Keep in mind that the training and testing phases are completely independent from each other. Training and model selection occurs on *in silico* generated data derived from online databases, while testing takes place on the experimentally acquired MS data for which we know the content.

3.1. *Mass defect as a rule-of-thumb to detect lipids*

It has been previously observed that signals belonging to a specific biomolecular class are more likely to be found in a particular fractional mass interval [25–27]. This phenomenon is often used to discriminate between peptides and other molecular species, for example by examining the mass defect of an ion peak, which for discrimination purposes is computationally equivalent to examining the fractional part of an exact mass value as is the case in Kendrick maps. Similarly, lipid molecules also have well-determined ranges comprising their masses and the fractional parts of these masses. An example of solely using the mass for discrimination is demonstrated in Fig. 2, which looks at the lipid and peptide content from our *in silico* data set within a mass range of 740–750 Da. The three histograms A–C show the number of lipids and peptides (or rather their protonated mono-isotopic ions), grouped into bins of respectively 1, 0.1, and 0.01 Da wide. It is clear from these histograms that if the mass resolution is relatively coarse (Panel A), all bins will report the presence of both lipid and peptide species. This observation illustrates that with a resolution of one Dalton in this mass range, it is impossible to separate the two molecule classes purely on the basis of mass. However, as the mass resolution grows finer, we start to see that many bins start reporting primarily (Panel B) or even exclusively (Panel C) a single class of molecules. This observation means that, if mass resolution allows, lipid and peptide species start occupying specific sub-areas of the mass domain and discrimination on the basis of mass alone becomes possible. The three different bin widths can be considered to represent mass spectrometers with different resolutions. The figure also shows the overall (monoisotopic) mass distributions of both molecule classes across the entire mass range considered in this study (Panel D). Although Fig. 2(A)–(C) looks at mass rather than the fractional part of mass, the general observation regarding mass resolution is in line with the use of mass defect rules-of-thumb in high-resolution measurements for quick and early interpretation (e.g. in FTICR measurements) [24]. Using only one of the possible types of information that can be extracted from the spectrum, namely ion mass, Fig. 2(A)–(C) demonstrates that the ability to discriminate between molecular classes is a function of both theoretically differentiating aspects and the practical ability of the instrument to capture that differentiating aspect. Encouraged by this observation, the following section extends the search for discriminating features beyond molecular mass alone and starts considering multiple features simultaneously, all of which can be extracted from an experimental mass spectrum.

3.2. *Training of the peptide-vs.-lipid classifier and assessment of feature sensitivity*

The previous section illustrates that the best discriminating features between lipids and peptides depend on the capabilities of the instrument in question. Important instrument parameters include mass accuracy, mass resolution, and the ability of the instrument to accurately measure isotope intensities. To evaluate the performance of the lipid-vs.-peptide classifier with respect to noisy data, 30 (6×5) *in silico* spectra sets were generated that introduce an error to the intensity and mass values of the theoretical isotope distribution. For each set, a lipid-vs.-peptide classifier is trained separately. The RF classifiers are evaluated by means of out-of-bag error estimates, which are a machine learning technique to retrieve unbiased estimates of the misclassification rate of a classifier. This misclassification rate can be regarded as the chance that a lipid or peptide will be classified incorrectly. As such, a misclassification rate of 0% is optimal. Table 2 presents the misclassification rates of the 30 classifiers in function of noise on the mass values (columns) and peak intensities (rows). Each cell in the table is the mean misclassification rate over five randomly selected subsets of peptides, with the standard deviation shown in parentheses.

When no error is added to the theoretical isotope distribution, a misclassification rate of 0.15% is achieved. As the resolution is reduced and noise is added, misclassification rates increase. For example, with a normally distributed error on the isotope peak intensities of $\sigma = 0.3$ and rounding the masses to the first decimal digit, a misclassification rate of 10.9% is obtained. The results show that in terms of misclassification rates, the influence of intensity noise is limited when mass values are accurate up to the 4th decimal digit. However, when the mass resolution deteriorates further, the misclassification rate grows fast as intensity noise increases. Table 2 indicates that lipid versus peptide classification is certainly feasible albeit with differing success rates depending on the instrumental capabilities. In order to gain more insight into the isotopic features responsible for successful differentiation, an assessment of the feature sensitivity is required, certainly since the importance of the isotope features depends on instrument capabilities. It is particularly useful to examine which features are crucial in the classification procedure and which are less so, such that this information can be

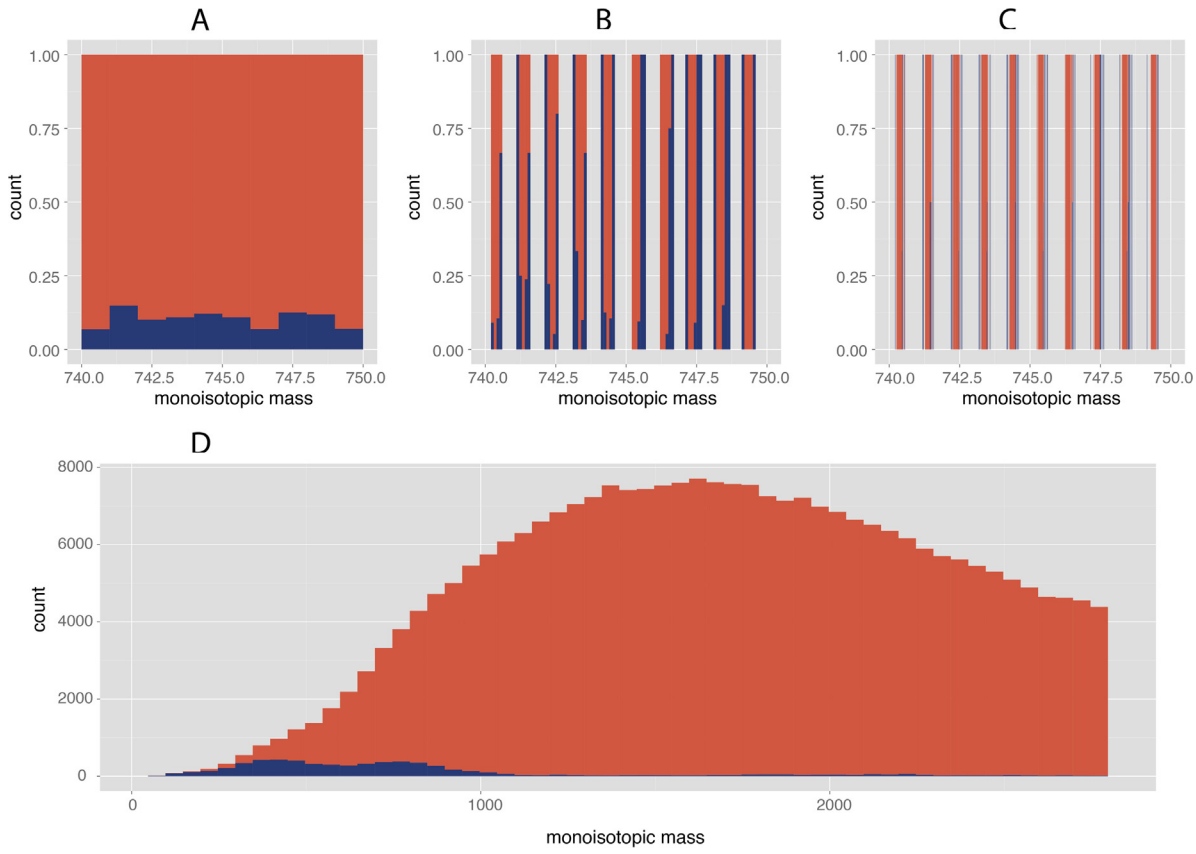

Fig. 2 – (A–C) Lipid and peptide counts between 740 and 750 Da. These histograms show the number of lipids (blue) and peptides (red) found (and normalized to 1) within a certain segment of mass range. From left to right, the mass resolution increases and the bin width narrows, corresponding to 1, 0.1, and 0.01 Da for panels A, B, and C, respectively. At coarse resolution (A), bins contain both lipids and peptides and mass-based discrimination is not feasible. At finer resolutions (B, C), bins start containing lipids or peptides exclusively, indicating that discrimination becomes possible. (D) Monoisotopic masses of all lipids (blue) and all *in silico* digested peptides (red) that were included in this study. The histogram shows the distribution of peptides on top of the distribution of lipids. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2 – Misclassification (out-of-bag) errors (in %) for two feature sets (complete and reduced) for different intensity noise levels and at different mass resolutions. These results cover RF classifiers that aggregate votes from 1000 trees. The mass resolution goes from theoretically perfect (left-most column) to 0.1 Da (right-most column). Intensity noise increases from top to bottom. Lower values are better and zero is optimal. The results show the average (and sd in parentheses) over 5 repetitions for sampling the training subset of peptides. The approximate relative mass resolution (in ppm) tied to the decimal digit rounding is included as top row. This approximation is based on the mass range of the molecules included in the study, where the lightest mass equals 47.01 Da, and the heaviest mass equals 2799.8 Da.

Approx. relative error for each column:	0 ppm	0.002–0.1 ppm	0.02–1.1 ppm	0.2–10.6 ppm	1.8–106.4 ppm	17.9–1063.5 ppm
Sd of intensity noise	No mass rounding	Mass rounding to 5th decimal digit	Mass rounding to 4th decimal digit	Mass rounding to 3rd decimal digit	Mass rounding to 2nd decimal digit	Mass rounding to 1st decimal digit
Complete feature set						
0	**0.15** (0.02)	**0.14** (0.03)	**0.18** (0.02)	**2.81** (0.08)	**4.34** (0.18)	**5.18** (0.23)
0.01	0.15 (0.03)	0.15 (0.02)	0.18 (0.03)	3.34 (0.14)	5.45 (0.15)	6.24 (0.28)
0.1	0.19 (0.03)	0.17 (0.01)	0.20 (0.03)	5.44 (0.09)	9.25 (0.36)	10.20 (0.34)
0.2	0.18 (0.03)	0.19 (0.01)	0.19 (0.03)	5.68 (0.11)	9.74 (0.31)	10.69 (0.33)
0.3	0.19 (0.03)	0.17 (0.01)	0.19 (0.02)	5.66 (0.14)	9.78 (0.29)	10.90 (0.31)
Reduced feature set						
	0.16 (0.027)	0.14 (0.015)	0.19 (0.012)	5.57 (0.136)	10.27 (0.361)	10..78 (0.129)

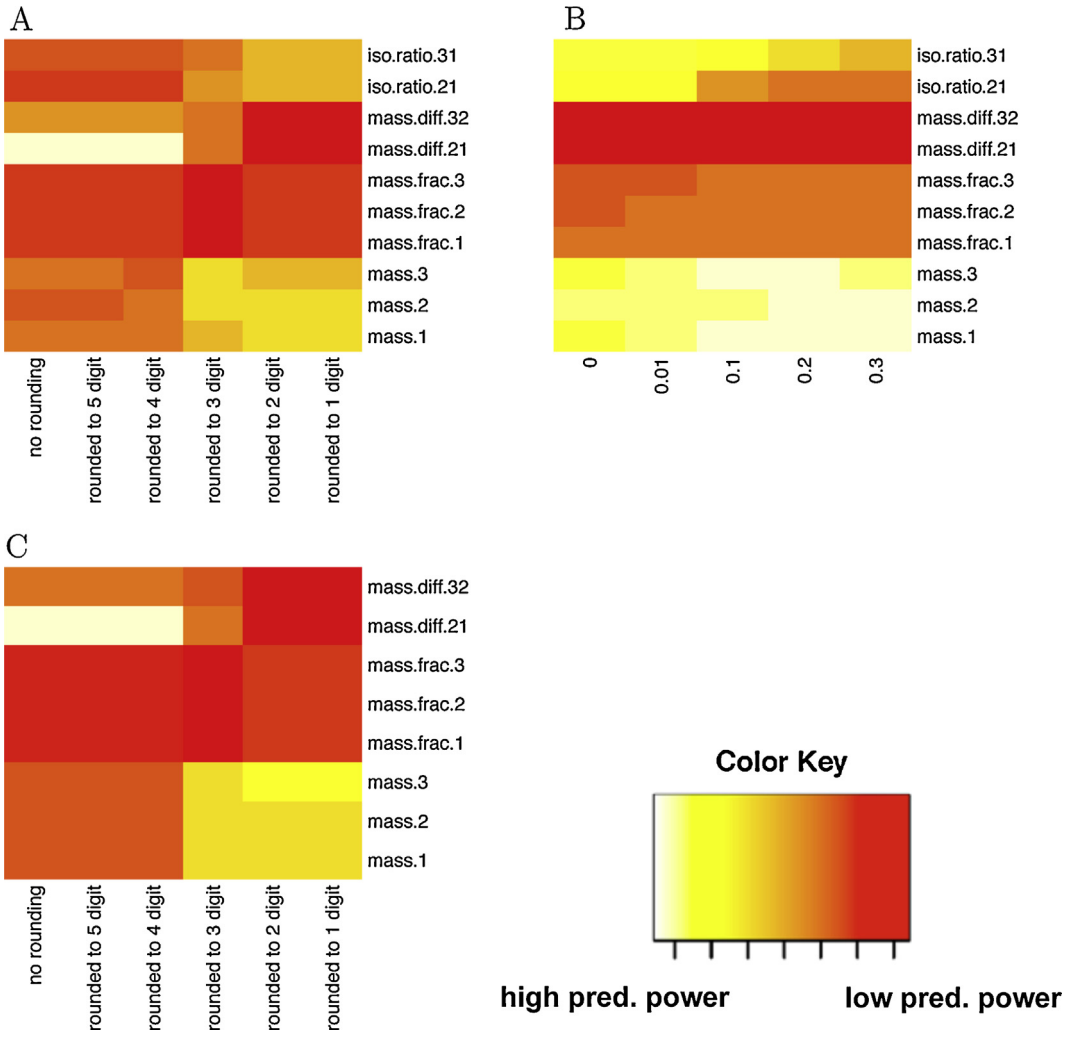

Fig. 3 – Predictive power of the isotopic features based on relative mean decrease in Gini index (abbreviations defined in Section 2). The red to white color scale corresponds to increasing feature importance (the same value might be represented by different colors in different panels). (A) Feature importance in function of mass resolution. For very accurate data (exact up to 3rd–4th decimal place), the features based on the mass difference between consecutive isotope peaks, *mass.diff21* and *mass.diff.32*, are important. When mass accuracy decreases, the importance of these features tends to decrease as well. In this plot the intensity features are without noise. (B) Feature importance as a function of noise on the intensity features ($\sigma = 0, \ldots, 0.3$) for masses rounded to the 2nd decimal digit. The addition of noise obscures the information content of the intensity features resulting into a decreasing importance. (C) Feature importance in function of the mass resolution (classifiers trained only on the reduced set of features). We observe trends similar to those in Panel A. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

incorporated into the design of future experiments. The importance of a feature in the classification process can be assessed through its Gini index. This metric captures the dispersion, or equivalently, the inequality caused by a split in the regression tree of a RF classifier. A Gini index equal to one represents a perfect separation and thus a value close to one corresponds to a feature with high predictive power. In order to obtain insight into the predictive power, and thus importance, of each feature within a particular resolution and noise context, we calculated the mean decrease in the Gini indices across the different spectra sets. The results are shown in Fig. 3. Panel A highlights the predictive power of each feature in function of decreasing mass resolution, with no noise added to the peak intensities (corresponding to values indicated in bold in Table 2). The lighter colors indicate higher predictive power and a higher importance connected to the feature in question. The darker the color, the lower its contribution to distinguishing between lipids and peptides. Note that the values in Fig. 3 are not the Gini index itself, but rather the relative (i.e. scaled by column) mean decrease in the Gini index.

Each column of Panel A corresponds to the best lipid-vs.-peptide classifier in that particular mass resolution case ranging from infinite mass resolution to 0.1 Da. Each column shows the relative importance of one particular feature versus the other features as the mass resolution changes. Panel A clearly shows that as long as the third decimal digit of the

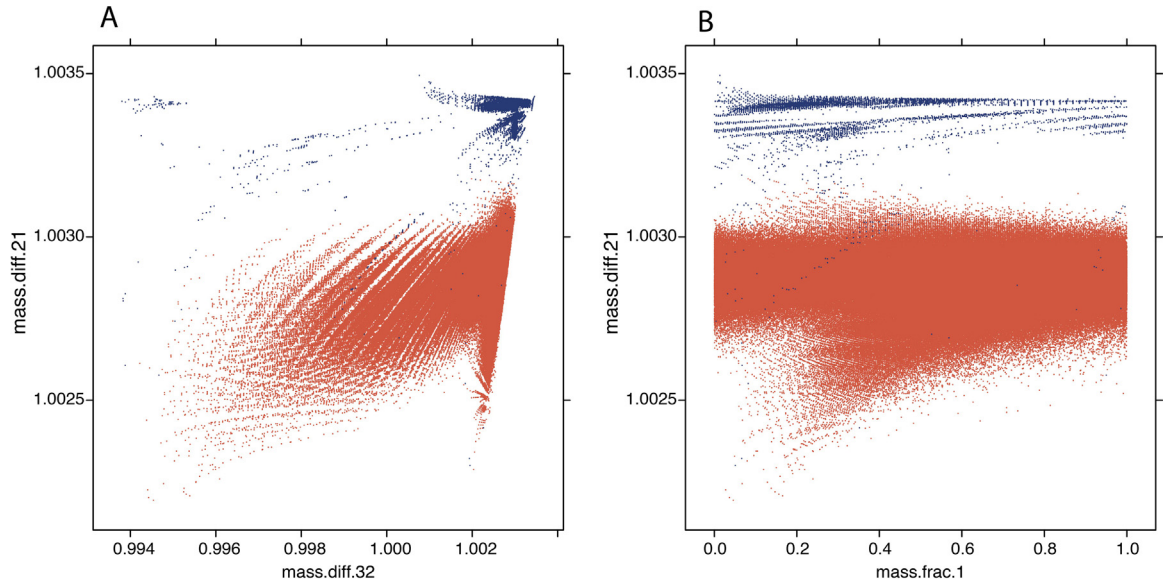

Fig. 4 – Distribution of the lipid (blue) and peptide (red) species along the *mass.diff.21* (difference between second and first isotope masses), *mass.diff.32* (difference between third and second isotope masses), and *mass.frac.1* (fractional part of mass of first isotope peak) dimensions in the case of infinite mass resolution. Note a clear separation between the two classes using these three features. As indicated in Fig. 3A, features *mass.diff.21* and *mass.diff.32* show the highest mean decrease in the Gini index for the infinite mass resolution case, further confirming these observations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mass values is accurate, the same features hold importance as in the theoretical infinite-resolution case. In other words, one could say that beyond 3 decimal digits in the mass values, the same features (and similar classifiers) can be used to distinguish between lipids and peptides. The most important features in this case are the mass differences between the first and second isotope peaks and the second and third isotope peaks. Note that although they contribute in this setting, ion intensity features (*iso.ratio.21* and *iso.ratio.31*) are much less telling than accurate mass determination features, which means that instrument sensitivity matters little in making a peptide/lipid distinction once a certain limit of detection is reached and three decimal digits are available for the masses. Also note that at these mass resolutions, the fractional mass features (e.g. the mass defect rule) are less informative than the mass differences between the first three isotopic peaks. This result seems to indicate that there might be other rules-of-thumb more informative than the mass defect if 3 or more decimal digits are available. There seems to be a tipping point in feature importance between masses rounded to 3 or more decimal digits and masses rounded to 2 or less decimal digits. At these lower mass resolutions, importance of *mass.diff.21* and *mass.diff.32* plummets as the isotope distribution becomes less and less well-described. As a result, the other features become more important for distinguishing between lipids and peptides. A logical explanation for this observation is that the discriminatory power of features *mass.diff.32* and *mass.diff.21* is driven by the carbon composition of the molecular species and, more specifically, the mass difference of 1.003 Da between the carbon isotopes ^{12}C and ^{13}C. At mass resolutions that are lower than three decimal digits this difference can be lost and other features will have to take over.

A different perspective on the change in importance ranking when rounding the mass is provided in Fig. 4. This figure shows the distribution of the lipid and peptide species in the training set along the *mass.diff.21*, *mass.diff.32*, and *mass.frac.1* dimensions (assuming infinite mass resolution). The entries from the Lipid Maps database are denoted by blue dots, while the *in silico* tryptic peptides are indicated by red dots. Note that in a theoretically infinite mass resolution setting a quasi-perfect separation between lipids and peptides is possible using just these three features. It is also clear that rounding masses to the second or first decimal digits would collapse the lipid and peptide data clouds and would obscure the differentiating information of the features based on mass differences between the isotopes.

Panel B of Fig. 3 highlights the predictive power of each feature in function of intensity noise, with the masses rounded to two decimal digits (corresponding to values indicated in italic in Table 2). Where Panel A suggests that the importance of the intensity features becomes substantial when masses with two or less decimal digits are available, Panel B evaluates the intensity features with respect to their sensitivity to increasing amounts of noise on the peak intensities (standard deviations ranging from $\sigma = 0$ to 0.3). It is clear that the importance of these intensity features diminishes as the noise increases. A similar observation was made for the data set rounded to the first decimal digit (data not shown). Mass spectral measurements are commonly corrupted by a substantial amount of error in the peak intensities. This noise usually originates from numerous latent variables and instrument artifacts that are hard to characterize. Another nuisance that affects the peak intensities in an isotope distribution is that saturated lipid species can overlap with their unsaturated variants. These

saturated isomers cause a significant bias in the observed isotope pattern. In light of these issues with ion intensity-derived features, we constructed a reduced feature set without the *iso.ratio.31* and *iso.ratio.21* features.

Another argument for restricting a classifier to features based on mass information alone is that mass resolution and accuracy can generally be better controlled via calibration and internal reference standards than ionization, analyzer and ion detector response.

In Panel C of Fig. 3 we highlight the predictive power of each feature in function of decreasing mass resolution, but this time only using the mass-derived features present in the reduced feature set (see the bottom row in Table 2). It is essential to understand that Panel C is not simply Panel A with the first two rows removed, but that the values in Panel C are the result of classifiers being trained only on the reduced set of features. The importance of the features in function of the different mass resolutions seems unaffected.

Given the independence from ion intensity that the reduced feature set brings and the relatively minor contribution of ion intensity-derived features, we will use one of these reduced set classifiers to test classification performance on a MS measurement (see Section 3.4). As we are dealing with a MALDI-TOF acquisition we feel confident that the masses are accurate up to the second decimal digit, which is why we will specifically apply the classifier trained on theoretical data rounded to the second decimal digit. As indicated in Table 2 this classifier demonstrates a misclassification rate of 10.27% on the theoretical training data. Section 3.4 will discuss its performance on the real MS data test set.

Overall, the misclassification rates of Table 2 seem to indicate that on the basis of our training set, a classifier can discriminate well between lipids and peptides. However, the question remains whether the training set is representative for undigested or bioactive peptides. It should be noted that bioactive peptides may have different characteristics than tryptic peptides, which always have an arginine or lysine at their C-terminus. To test the performance of the classifier on bioactive peptides, we perform a classification on virtual spectra generated from the undigested Uniprot entries lighter than 2.8 kDa. Analogous trends as observed in Table 2 are obtained for the undigested peptides (cf. Supplementary Table S2). Presumably, the structure of the amino acid chains conserves the characteristics for differentiation for undigested peptides as well.

3.3. *Multi-class classifier training*

This section elaborates on the sub-categorization of an isotope pattern after it has been annotated as a lipid. Fig. 5 displays the projection of the theoretical isotope features from the Lipid Maps database onto a coordinate system with the mono-isotopic mass as abscissa and its fractional part on the ordinate (in the case of infinite mass accuracy). The plot reveals a visible separation between some of the lipid classes as defined by the Lipid Maps Consortium. Unfortunately, for some classes a large overlap between the different lipid species is apparent, which makes it difficult to accurately pinpoint the subclass via a classification strategy. However, information from the plot could still be used to enrich the

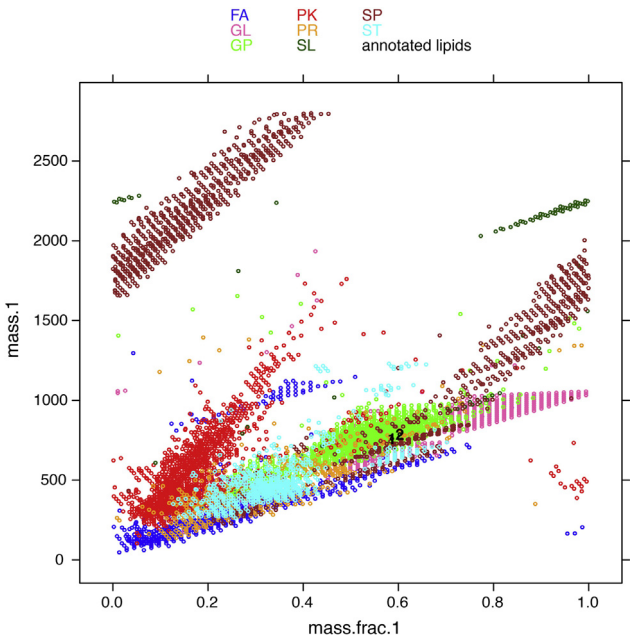

Fig. 5 – Distribution of lipid subclasses from the training set, along the *mass.1* (exact mass of first isotope peak) and *mass.frac.1* (fractional part of mass of first isotope peak) dimensions for infinite mass resolution. Black labels 1 and 2 (fractional mass: 0.58 and 0.6/mass: 760.58 Da and 786.60 Da) indicate two glycerophospholipids correctly classified via a random forest classifier (cf. Table 4). Note that a random forest classifier operates on a multidimensional feature space, which contains more information than this two-dimensional visual map can represent. The features used in this figure were chosen for illustrative purposes.

probability of correct annotation for at least some of the lipid subclasses. For example, a lipid molecule with a mono-isotopic mass of ≈800 Da and a fractional mass of ≈0.2 Da has a high likelihood of being a polyketide (PK) and not a glycolipid (GL). On the other hand, a lipid molecule with a mono-isotopic mass of ≈800 Da and a fractional mass of ≈0.8 Da is likely to be a glycolipid and much less likely to be a polyketide. This plot further extends the idea of visual maps by Hughey et al. [28] for within-lipids classification. Also note that the RF classifier can employ many more features than the two employed in Fig. 5. For the lipid sub-categorization task we use the reduced feature set and train a multi-class random forest classifier for the 8 lipid classes. In order to be compatible with the testing phase on a real MALDI-TOF spectrum, the RF model was trained on a data set that reflects a mass resolution of TOF class instruments, with mass values rounded to the second decimal digit. The misclassification rate for the multi-class classifier is presented in the confusion matrix of Table 3. Although the overall misclassification error is large (over 30%), some lipid classes are better discernible than others. A case in point are the classes GP, PK, and SP, which have a misclassification rate less than 17%. From the visual map in Fig. 5 it can be seen that the classes PK and SP are well separable from the other lipid classes. The class GP is harder to distinguish, but it contains

Table 3 – Misclassification rates per lipid subclass by the model trained on the reduced feature set and *in silico* data that is rounded to the second decimal digit. The row label indicates the correct identity, the column label indicates the identity suggested by the multi-class classifier. For three large subclasses – GP, PK, SP – with more than 1150 lipids in each of them, the misclassification (out-of-bag) error is smaller than 17%. The RF classifier was built using 1000 trees.

	FA	GL	GP	PK	PR	SL	SP	ST	$\sum_{row}$	Class. error (%)
FA	531	16	24	60	102	0	31	149	913	41.8
GL	14	255	57	6	8	1	30	29	400	36.2
GP	11	48	1178	36	24	2	47	69	1415	16.8
PK	38	1	51	1133	29	0	4	40	1296	12.6
PR	151	10	69	67	44	2	13	86	442	90
SL	0	7	3	3	1	49	12	1	76	35.5
SP	54	26	82	3	7	3	974	18	1167	16.5
ST	150	31	101	49	67	0	5	201	604	66.7
$\sum_{column}$	949	394	1565	1357	282	57	1116	593	6313	Total class. err: 30.9

regions that are well discernible. Note that these three classes cover more than 60% of the lipids in the Lipid Map database, raising issues of balance but at the same time suggesting that for the majority of lipids the sub-categorization is theoretically feasible based on mass properties.

3.4. *Classifier test on real MS data*

While building the lipid-vs.-peptide and multi-class classifiers, several decisions are made that influence the operation of these classifiers. Some of these decisions may have a negative impact on the performance of the classifier. For example, one can question whether the mass accuracy of two decimal digits to represent TOF class instruments is too loose or rather too conservative. In fact, this parameter mainly depends on the particularities of the instrument and how it has been calibrated, so it is not entirely correct to assume a general parameter for TOF class instruments. Further, based on our previous experiences with small molecules, we choose to work with isotope distributions which contain at least three isotopic peaks. This restriction can either be considered too stringent or too lenient. Another constraint is that the training phase does not consider modifications or adduct formations, which lipids and peptides can undergo in real experiments. In this paper we have only considered the protonation of an ion. In order to assess whether our assumptions hold up in a genuine measurement, we test the classifiers on real MS data. For this purpose, a controlled experiment is conducted in which a mixture of five peptide and seven lipid species is assayed through MALDI-TOF MS. This experiment delivers a mass spectrum from which we retain only the ions for which the first three isotope peaks can be found within an interval of ±0.01 Da around their theoretical (protonated) mass. Six isotope families are found that meet this criterion and they are summarized in Table 4 with their presumed identity as provided by the key in Table 1. It should be mentioned that a relaxed search to an interval of ±0.1 Da did not result in additional findings. After extracting from the mass spectrum the values of the reduced set of features for each of these isotope families, the lipid-vs.-peptide and multi-class classifiers use the measured features to predict the identity of the compound.

The classifier returns a probability close to one for peptides and probabilities close to zero to indicate lipids. For the six considered molecules the average probabilities for lipids no. 1 and 2 were 0.0126, 0.0004, respectively, and peptide no. 3, 4, 5, 6 received a probability of 0.1576, 0.9738, 0.9996, 0.8476, respectively (σ for molecule 3: 0.189; other $\sigma < 0.017$). Thus it seems that molecules no. 1 and 2 are correctly classified as lipids. Also molecules no. 4, 5 and 6 are correctly annotated as peptides. Overall, this is a strong result that seems to indicate our assumptions are valid and that automated annotation of mass spectra without supervision is feasible. However, note that peptide no. 3 has been misclassified, which points to uncertainty regarding the molecular class that has generated the observed signal. In order to gain more insight into the origin of this confusion, we visualize the experimental and *in silico* data with respect to the mono-isotopic peak mass and its fractional part. The result is displayed in Fig. 6 and extends the concept of the visual maps described in the section about multi-class prediction. In this figure, peptides are represented by purple dots and lipids are indicated by blue dots. Note that lipids exhibit more heterogeneity in their distribution than peptides. It can be ascertained that peptide no. 3 is indeed positioned near the border of the subspace that is occupied by lipids. Its border position may explain the weak probability returned by the lipid-vs.-peptide classifier. The other molecules are more at a center position in the data cloud of their respective molecular classes, which is also reflected by the strong probabilities close to zero and one. However, the figure only presents a projection of the data in a reduced space and should not be over-interpreted, since the RF classifier involves more parameters to support its decision. For weak probabilities, i.e. near 50%, we could consider introducing an additional category that collects uncertain assignments by applying an upper and lower limit on the RF probabilities. As a whole these test set results seem to indicate that training a classification model on *in silico* data is justified for an application on mass spectrometry measurements. The independence between the theoretical training data set and the real MS test data set ensures that the machine learning classifier does not overfit the model on measurement-specific features in the data, which has happened in former classification studies [29,30] as pointed out by Baggerly et al. [31].

The two observed isotope patterns that were correctly classified as lipids were successfully assessed via the multi-class RF classifier as well. The signals corresponding to lipid PC(16:0/18:1(9z)) and PC 18:1(9z)/18:1(9z) were correctly annotated as glycerophospholipids (GP), with a classifier certainty of 98% and 82.7%, respectively. This result is not surprising given that the GP class can be quite well discriminated

Table 4 – Molecules annotated in the mass spectral measurement. m_o = observed values of three first center-masses starting from the mono-isotopic peak; iso_o = observed isotopic abundances for first three aggregated peaks; $10^6 \times ((m_t - m_o)/m_o)$ = differences between theoretical center-masses obtained by BRAIN and observed MS data (in ppm); $iso_t - iso_o$ = differences between theoretical isotopic abundances of first three isotopic peaks obtained by BRAIN and observed MS data.

No.	Name	Type	Formula	m_o			$10^6 \times \frac{m_t - m_o}{m_o}$			iso_o		$iso_t - iso_o$	
				mass.1	mass.2	mass.3				iso.ratio.21	iso.ratio.31		
1	PC 18:1(9z)/16:0	Lipid	$C_{42}H_{83}N_1O_8P_1$	760.5864	761.5899	762.6012	−1.0176	−1.2198	−12.0718	0.4383	0.5406	0.0323	−0.41589
2	PC 18:1(9z)/18:1(9z)	Lipid	$C_{44}H_{85}N_1O_8P_1$	786.6029	787.6059	788.6091	−2.0544	−1.6214	−1.7385	0.4234	0.0958	0.0690	0.03930
3	PKC substrate	Peptide	$C_{34}H_{69}N_{16}O_8$	829.5501	830.5521	831.5536	−2.07462	−1.29673	−0.23811	0.4637	0.1379	−0.0259	−0.02767
4	ACTH 4-11	Peptide	$C_{50}H_{72}N_{15}O_{11}S_1$	1090.5279	1091.5308	1092.5339	−2.0880	−2.1639	−4.1225	0.5795	0.2615	0.0372	−0.00658
5	Glu1 Fibrinopeptide B	Peptide	$C_{66}H_{96}N_{19}O_{26}$	1570.6810	1571.6838	1572.6852	−2.3047	−2.2854	−1.4752	0.8284	0.3794	−0.0235	−0.00599
6	ACTH 18-39	Peptide	$C_{112}H_{166}N_{27}O_{36}$	2465.2065	2466.2102	2467.2114	−3.1081	−3.3878	−2.7322	1.3070	0.9737	0.0369	−0.00342
Additionally annotated molecules													
7	PC 18:1(9z)/18:1(9z) or PC 18:1(6z)/18:1(6z) or PC 18:1(9trans) with Na^+	[Lipid]	$C_{44}H_{84}N_1O_8P_1Na_1$	808.5849	809.5880	810.5908	−2.0777	−1.7169	−1.4064	0.4676	0.1253	0.0247	0.00977
8	Glu1 Fibrinopeptide B w/o water	[Peptide]	$C_{66}H_{94}N_{19}O_{25}$	1552.6705	1553.6702	1554.6744	−2.37011	−0.34756	−1.33147	0.7564	0.4200	0.0480	−0.04908

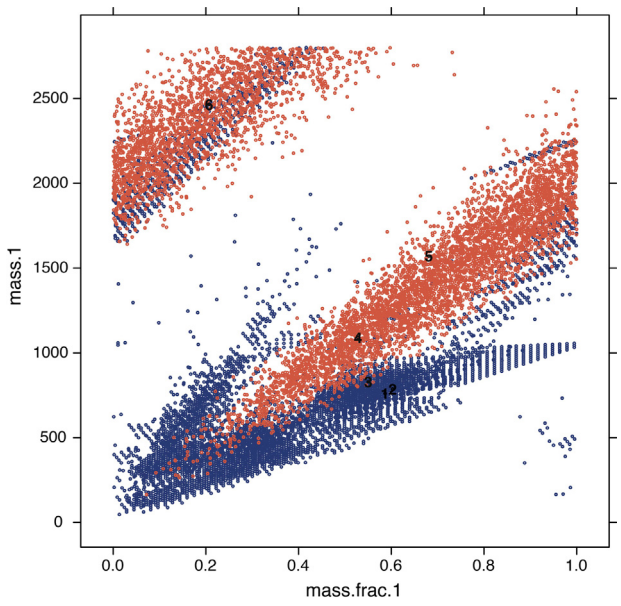

Fig. 6 – Distribution of lipids (blue) and a random subset of *in silico* digested peptides (red) along the *mass.1* (exact mass of first isotope peak) and *mass.frac.1* (fractional part of mass of first isotope peak) dimensions for infinite mass resolution. In addition the annotations of molecules measured in the MS experiment were marked with black numeric labels (cf. Table 4). The features used in this figure were chosen for illustrative purposes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in theory, as revealed by the relatively low misclassification rate of 16.8% in Table 3. The visual map of Fig. 5 projects the two lipids into the *mass.1* and *mass.frac.1* dimensions. Although some overlapping lipid classes occur for the GP class, the two considered lipid molecules are clearly positioned within the GP data cluster and are differentiated from the other possible lipid classes.

Since the spectrum contains more than just the protonated ions expected from the mixture, an automated peak extraction and subsequent classification is performed on all the found isotope patterns. A post-hoc analysis of the results reveals that two of the annotated isotope distributions can be traced back to modified forms of the standards in the mixture. For example, one of the patterns corresponding to either PC 18:1(9z)/18:1(9z), PC 18:1(6z)/18:1(6z) or PC 18:1(9trans) with a sodium adduct was classified as a lipid (probability of being a peptide 0.0074; $\sigma < 0.012$). Another isotope pattern is annotated as a peptide with an average probability of 0.999 ($\sigma < 0.0015$) and can be speculated to be a protonated Glu1 Fibrinopeptide B which has lost a water molecule. The isotope patterns for the two molecules correspond to the theoretical distribution within an interval of 2.4 ppm for the mass and a margin of 0.05 on the normalized intensities as indicated in the second part of Table 4. These findings suggest that the lipid-vs.-peptide classifier that is trained only on protonated ions might generalize to other modifications and adducts as well.

Although the method is evaluated on a MALDI spectrum, it should work equally well on data obtained via other ionization sources (e.g. ESI) since the isotope information is a characteristic of the molecule and not a characteristic of the ionization process. In the case of ESI, the added complexity of multiple charging might require a spectrum to undergo charge-state deconvolution first though.

4. Discussion and concluding remarks

In this manuscript we present a proof-of-concept study that evaluates whether machine learning methods can be employed to aid in the automated interpretation of full scan mass spectra. The first objective in this study was to find which isotope features, on a theoretical basis at infinite mass resolution, can function as differentiating criteria between peptides and a generic lipid class. We found that the fractional mass and the mass differences between isotopic variants are important to drive that classification. The high discriminatory power of these features can be explained by the different proportion of carbon present in lipids and peptides [24,32]. An *in silico* analysis confirmed some of the empirical rules that are common practice in lipidomic filtering (e.g. mass defect). The second objective of the study was to see whether these theoretical findings are practical in mass spectral measurements. In other words, what happens if mass resolution and spectral sensitivity is finite? We conducted an *in silico* sensitivity analysis to assess the robustness of isotopic features at different resolutions and noise levels in accordance with the different instruments on the market today. It is interesting to see that mass-difference-between-isotopic-peaks features drop from highly important to a low importance when the mass resolution of the instrument deteriorates. At the same time, the isotope intensities gain importance for instruments with a lower mass resolution. This transition is sensible since a loss of information in the masses is compensated by information from isotopic peak intensities. However, peak intensities are usually affected by a severe amount of noise, which often is difficult to characterize since it depends on multiple unobserved instrument factors. To make the random forest classifier robust to a misspecification in the noise structure, we motivate to exclude information about peak intensities from the model because there are ample noise sources that can severely disturb the measured isotope profile. Nevertheless, as instrumentation evolves and spectral accuracy improves, isotopic peak intensities could again be incorporated into the decision making process of the random forest classifier.

The third objective was to determine whether the isotope features can be used to sub-categorize the lipid classes once a molecular ion is classified as lipid. Here, similar conclusions as with the lipid-vs.-peptide classifier can be drawn. Perfect classification of the eight lipid classes is not achieved during this sensitivity study, however, enrichment of the probability that a particular lipid belongs to certain classes is attainable. Further, it should be noted that it is very unlikely that the eight lipid classes are simultaneously present in the data. It is also important to note that instrumental constraints such as ionization efficiency and limit of detection are not part of the classification study. The advantage of molecule-driven features rather than instrumentation ones is that instrumental argumentation does not contaminate the classification rules learned here. This approach ensures that the conclusions hold true regardless of how instrumentation develops in the future.

The fourth objective was to test on a concise MALDI-TOF MS experiment selected lipid-vs.-peptide and multi-class classifiers, trained solely on computer-generated data. This test is essential in our investigation as it provides proof that assumptions regarding the model are justified. For this purpose, we select the classifier that was trained for the classifiers that were trained for a mass resolution up to two decimal digits. Six isotopic profiles of known molecules were fed into the RF classifier of which five were correctly annotated. The probability returned by the RF classifier can be used to score the decision strength, with a score near zero or one indicating a clear separation between the two classes. One molecule received a class probability close to 16% indicating misclassification on behalf of the classifier regarding this isotope pattern. This result is probably caused by the peptide exhibiting features that put it close to the plane of separation between the two classes. Currently, two class labels are included in the model, but one option is to include a third no-lipid/no-peptide class that collects data of unknown molecular class. The two isotope patterns that were annotated as a lipid were further categorized by the multi-class classifier and correctly recognized as glycerophospholipids.

Although this study is not exhaustive, it does demonstrate one type of framework within which one can enable automatic interpretation of empirically acquired mass spectra. The approach is limited only by the quality of the databases that we provide for the species of interest and the practical feasibility of parameters in the machine learning process. We thus demonstrate one implementation but we recognize that this is not an exhaustive treatment and further improvements are possible with more advanced computational resources and more elaborate techniques (e.g. support vector machines). However, the proof-of-concept we provide for automated annotation of mass spectra delivers encouraging and useful results with relative little resources.

Overall, this paper demonstrates through a theoretical assessment and an empirical test, a theoretical basis for automated interpretation of lipids versus peptides and lipid class sub-categorization. Because of this ability, we propose to name the method *Lipid Centrifuge* for further references. This approach can also be potentially extended towards other molecule classes or towards further subdivision, for example, the detection of glycosylated peptides or adducts, such as acetate, ammonium, and formiate. On the other hand, this might not be absolutely necessary since our model, which is built without taking adducts into account, is shown to correctly classify some adduct species already. This indicates that some of these rules could be sufficiently general to cover adducts as well.

Given the extent of applications for automated interpretation of full scan spectra (e.g. imaging mass spectrometry, on-the-fly determination of the optimal MS/MS fragmentation strategy, selection of a downstream identification analysis path, etc.), we are convinced that this line of research merits further development.

Acknowledgements

This research is supported in part by the Polish National Science Center grant 2011/01/B/NZ2/00864 and by the EU through the European Social Fund, contract number UDAPOKL. 04.01.01-00-072/09-00. A.G, D.V. and P.D. gratefully acknowledge the support of the bilateral FWO-PAS grant VS.005.13N/Innovative algorithms to detect protein modifications in mass spectrometry data. P.D. is supported by a START fellowship from the Foundation for Polish Science. K.L. and D.V. acknowledge the support of the SBO grant 'InSPECtor' (120025) of the Flemish agency for Innovation by Science and Technology (IWT). R.C., G.H., and R.V. acknowledge support by the National Institutes of Health via grants NIH/NIGMS R01 GM058008-14 and NIH/NIGMS P41 GM103391-03.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.euprot.2014.05.002.

REFERENCES

[1] Brown HA, Murphy RC. Working towards an exegesis for lipids in biology. Nat Chem Biol 2009;5:602–6.
[2] Stoeckli M, Chaurand P, Hallahan DE, Caprioli RM. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. Nat Med 2001;7:493–6.
[3] Van de Plas R, Ojeda F, Dewil M, Van Den Bosch L, De Moor B, Waelkens E. Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis. Pac Symp Biocomput 2007:458–69.
[4] Van de Plas R. Tissue Based Proteomics and Biomarker Discovery – Multivariate Data Mining Strategies for Mass Spectral Imaging [Ph.D. thesis]. Leuven, Belgium: Faculty of Engineering, K.U. Leuven; 2010.
[5] Kirchner M, Timm W, Fong P, Wangemann P, Steen H. Non-linear classification for on-the-fly fractional mass filtering and targeted precursor fragmentation in mass spectrometry experiments. Bioinformatics 2010;26:791–7.
[6] Bruce C, Shifman MA, Miller P, Gulcicek EE. Probabilistic enrichment of phosphopeptides by their mass defect. Anal Chem 2006;78:4374–82.
[7] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[8] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. 10th ed. New York: Springer-Verlag; 2013.
[9] Fahy E, Subramaniam S, Murphy RC, Nishijima M, Raetz CRH, Shimizu T, et al. Update of the lipid maps comprehensive classification system for lipids. J Lipid Res 2009;50(Suppl):S9–14.
[10] Claesen J, Dittwald P, Burzykowski T, Valkenborg D. An efficient method to calculate the aggregated isotopic distribution and exact Center-Masses. J Am Soc Mass Spectrom 2012;23(4):753–63.
[11] Valkenborg D, Mertens I, Lemière F, Witters F, Burzykowski T. The isotopic distribution conundrum. Mass Spectrom Rev 2012;31:96–109.
[12] Rockwood AL, Palmblad M. Isotopic distributions. Methods Mol Biol 2013;1007:65–99.
[13] Scheubert K, Hufsky F, Böcker S. Computational mass spectrometry for small molecules. J Cheminform 2013;5:12.
[14] Böcker S, Letzel MC, Lipták Z, Pervukhin A. Sirius: decomposing isotope patterns for metabolite identification. Bioinformatics 2009;25:218–24.
[15] Fernandez-de Cossio Diaz J, Fernandez-de Cossio J. Computation of isotopic peak center-mass distribution by Fourier transform. Anal Chem 2012;84:7052–6.
[16] Böcker S. Comment on: "An efficient method to calculate the aggregated isotopic distribution and exact center-masses" by Jürgen Claesen, Piotr Dittwald, Tomasz Burzykowski, Dirk Valkenborg, J. Am. Soc. Mass Spectrom. 23 (2012) 753–763. J Am Soc Mass Spectrom 2012;23:1826–7.
[17] Royal Society of Chemistry ChemSpider. The free chemical database. http://www.chemspider.com
[18] Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, et al. METLIN: a metabolite mass spectral database. Ther Drug Monit 2005;27:747–51.
[19] Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 2010;45:703–14.
[20] The UniProt Consortium. Reorganizing the protein space at the universal protein resource (UniProt). Nucleic Acids Res 2011;40(D1):D71–5.
[21] Dittwald P, Claesen J, Burzykowski T, Valkenborg D, Gambin A. BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. Anal Chem 2013;85:1991–4.
[22] Carvalho PC, Xu T, Han X, Cociorva D, Barbosa VC, Yates JR. YADA: a tool for taking the most out of high-resolution spectra. Bioinformatics 2009;25:2734–6.
[23] Breiman L. Random Forests manual. http://oz.berkeley.edu/breiman/RandomForests/cc_home.htm [accessed 29.07.13].
[24] Sleno L. The use of mass defect in modern mass spectrometry. J Mass Spectrom 2012;47:226–36.
[25] Egertson JD, Eng JK, Bereman MS, Hsieh EJ, Merrihew GE, MacCoss MJ. De novo correction of mass measurement error in low resolution tandem MS spectra for shotgun proteomics. J Am Soc Mass Spectrom 2012;23:2075–82.
[26] Mitra I, Nefedov AV, Brasier AR, Sadygov RG. Improved mass defect model for theoretical tryptic peptides. Anal Chem 2012;84:3026–32.
[27] Frahm JL, Howard BE, Heber S, Muddiman DC. Accessible proteomics space and its implications for peak capacity for zero-,one- and two-dimensional separations coupled with FT-ICR and TOF mass spectrometry. J Mass Spectrom 2006;41:281–8.
[28] Hughey CA, Hendrickson CL, Rodgers RP, Marshall AG, Qian K. Kendrick mass defect spectrum: a compact visual analysis for ultrahigh-resolution broadband mass spectra. Anal Chem 2001;73:4676–81.
[29] Petricoin III EF, Ardekani AI, Ali M, Hitt BA. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–7.
[30] Conrads TP, Fusaro VA, Ross S, Johann D, Rajapakse V, Hitt BA, et al. High resolution serum proteomic features for ovarian cancer detection. Endocr-Relat Cancer 2004;11:163–78.
[31] Baggerly KA, Morris JS, Coombes KR. Reproducibility of seldi-tof protein patterns in serum: comparing data sets from different experiments. Bioinformatics 2004;20:777–85.
[32] Nielsen ML, Savitski MM, Zubarev RA. Improving protein identification using complementary fragmentation techniques in Fourier transform mass spectrometry. Mol Cell Proteomics 2005;4:835–45.