

RESEARCH ARTICLE

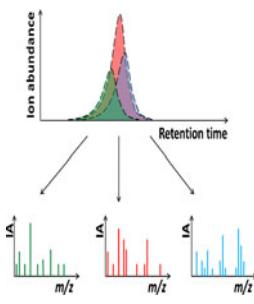
Clustering and Filtering Tandem Mass Spectra Acquired in Data-Independent Mode

Huisong Pak,¹ Frederic Nikitin,² Florent Gluck,^{1,3} Frederique Lisacek,² Alexander Scherl,^{1,3} Markus Muller^{1,2}

¹University of Geneva, Geneva, Switzerland

²SIB Swiss Institute of Bioinformatics, University Medical Center, 1, Rue Michel-Servet, 1211 Geneva 4, Switzerland

³Swiss Centre for Applied Human Toxicology, Geneva, Switzerland



Abstract. Data-independent mass spectrometry activates all ion species isolated within a given mass-to-charge window (m/z) regardless of their abundance. This acquisition strategy overcomes the traditional data-dependent ion selection boosting data reproducibility and sensitivity. However, several tandem mass (MS/MS) spectra of the same precursor ion are acquired during chromatographic elution resulting in large data redundancy. Also, the significant number of chimeric spectra and the absence of accurate precursor ion masses hamper peptide identification. Here, we describe an algorithm to preprocess data-independent MS/MS spectra by filtering out noise peaks and clustering the spectra according to both the chromatographic elution profiles and the spectral similarity. In addition,

we developed an approach to estimate the m/z value of precursor ions from clustered MS/MS spectra in order to improve database search performance. Data acquired using a small 3 m/z units precursor mass window and multiple injections to cover a m/z range of 400–1400 was processed with our algorithm. It showed an improvement in the number of both peptide and protein identifications by 8 % while reducing the number of submitted spectra by 18 % and the number of peaks by 55 %. We conclude that our clustering method is a valid approach for data analysis of these data-independent fragmentation spectra. The software including the source code is available for the scientific community.

Keywords: Proteomics, Mass spectrometry, Data clustering, Data-dependent, Data-independent, Shotgun proteomics, PACIFIC, SWATH

Received: 4 May 2013/Revised: 22 July 2013/Accepted: 29 July 2013/Published online: 5 September 2013

Introduction

Combination of orthogonal methods such as liquid chromatography (LC) and mass spectrometry (MS) is the main analytical system involved in proteomics. Complex mixtures of peptides are separated by reverse-phase (RP) LC and gas-phase molecular ions are formed during electrospray ionization (ESI) prior to MS detection. Peptides emitted by ESI are isolated, activated by collision induced dissociation (CID), and fragment ions are detected and analyzed for

peptide identification. Traditionally, a data-dependent acquisition (DDA) strategy is used for bottom-up proteomics. This method allows sequentially isolating and activating a number of most abundant precursor ions detected in a survey scan (MS1) prior to acquiring tandem mass spectra (MS/MS). To avoid redundant selection of the same peptides during chromatographic separation and to sample analytes more efficiently, peptides already selected are excluded for a given time (dynamic exclusion) after a first selection [1]. DDA data display a bias towards abundant peptides and, thus, show mainly abundant proteins of a proteome. The selection of low-abundance peptides is limited by the intra-spectrum dynamic range for MS1 spectra of the mass spectrometer. Also, low abundance peptides are often masked by high abundant ones and their selection for MS/MS is a rare event [2]. However if these low intensity peptides are isolated and fragmented, identifiable MS/MS spectra can be acquired [3]. The other observation related to

Electronic supplementary material The online version of this article (doi:10.1007/s13361-013-0720-z) contains supplementary material, which is available to authorized users.

Correspondence to: Markus Muller; e-mail: Markus.Mueller@ib-sib.ch

DDA is its poor reproducibility of peptide selection during LC separation, especially for low abundance peptides [2, 4]. Typically, even if some MS/MS spectra of high abundance peptides were repeatedly measured, differences in terms of retention time, intensity, and exclusion list are observed over several DDA of the same sample.

An alternative method to DDA would be an unbiased peptide selection for MS/MS during chromatographic separation of analytes regardless of their abundance. In 2003, Purvine and co-workers described a strategy called shotgun-CID, which is based on in-source (IS) fragmentation of precursor ions by using two different nozzle-skimmer voltage potentials in quadrupole time-of-flight (Q-ToF) instrument [5]. The main idea was to have sequential pairs of low and high collision energy spectra of peptides eluting from chromatographic separation without any isolation of precursor ions. The low energy spectrum contains mainly molecular ions, whereas the high energy spectrum contains mainly fragment ions from all present species. For data analysis, precursor and fragment ion chromatograms are extracted and elution patterns are correlated. The precursor and fragment ion lineage is reconstructed on this basis prior to database search, where a high-resolution mass spectrometer is recommended to obtain the required accuracy for precursor and fragment ions. The LC-MS^E method developed by Waters (Milford, MA, USA) is an example of the shotgun-CID technology in proteomics [6]. Another method for selecting precursor ions in an unbiased manner is to isolate and fragment all ion species within a given *m/z* window. Such methods are called data-independent acquisition (DIA) in contrast to DDA. MS/MS spectra acquired in the given *m/z* range of precursor ion isolation window are less complex compared with shotgun-CID spectra, due to the limited co-isolation. Venable and co-workers used relatively large but limited isolation windows for precursor ion isolation prior to MS/MS, typically 10 *m/z* units [7]. Because co-fragmentation and co-elution were a major issue with such large isolation windows software deconvolution based on chromatographic elution time had to be used prior to database search. Panchaud *et al.* introduced the PAcIFIC (Precursor Acquisition Independent From Ion Count) method [8] extending Venable's concept to smaller isolation windows, typically 3 *m/z* units. An instrument acquisition cycle consists typically of acquiring 10 to 25 consecutive MS/MS spectra, with an isolation window of about 3 *m/z* units and an increment of the center of the isolation window by 2 to 3 *m/z* units between two consecutive MS/MS spectra in a linear ion trap. A total *m/z* range of 15 to 50 units is thus covered during one LC-MS/MS cycle. The small isolation window is comparable to DDA. Precursor ion species within the isolation window are fragmented regardless of their abundance during their entire chromatographic elution. To cover the desired *m/z* range for a full experiment, the sample is repeatedly injected and during each injection a different precursor ion isolation range is used. The *m/z* range is typically 400–1400 for a proteomics experiment with trypsin

as cleavage enzyme. Thus, the sample needs to be injected 20 to 40 times, in a concept similar to gas-phase fractionation [9, 10]. Such an approach became feasible thanks to the tremendous progress in ion trap instrument acquisition frequency. In comparison to DDA, the method uses the same window for precursor ion isolation. Thus, the same ratio of peptide co-fragmentation resulting with chimeric mass spectra is observed. However, it increases drastically the dynamic range of peptide/protein identification. Values of 10⁷ across the chromatographic experiment are reported with current instruments [8]. Panchaud and co-workers have reported how the three essential parameters (i.e., number of MS/MS scans events per cycle, precursor isolation window width, and *m/z* channel increment affect the duty cycle and the analysis performance [11]). This approach was recently used for several applications in the field of proteomics. Chen and co-workers also reported the feasibility of combining PAcIFIC with direct infusion of samples into the mass spectrometer. According to their results, PAcIFIC can typically be used for medium complex samples within a fast analysis time (in the rate of a few minutes rather than hours and days) [12]. The latter demonstrates again the efficiency of “systematically interrogating all *m/z* channels for the presence of peptides regardless of the observation of precursor ions” [13].

Recently, Aebersold's group presented a strategy called SWATH MS [14]. It is a DIA strategy that fully exploits the advantages of DIA for peptide/protein quantification. It uses a window of 25 *m/z* to isolate precursor ion species and fragment them all in a Q-q-TOF. The particularity of SWATH resides in its acquisition of MS/MS spectra with high measured accuracy (10–50 ppm) and mass resolution (15,000–30,000) at a high scan rate (duty cycle of 3.3 s to acquire 32 MS/MS + 1 MS1). Consequently, one SWATH acquisition is sufficient to cover a mass range from 400 to 1200 *m/z* units suitable for proteomics applications. The frequency of MS/MS spectra acquisition allows sufficient data points for peptide quantification across each chromatographic peak. However, because of the large precursor isolation window used with SWATH, the acquired data are not optimal for direct identification using database search. Thus, DDA acquisition is usually required in a first step for peptide/protein identification. These identifications are collected to build a database of precursor-fragment ion transitions and exploited for subsequent quantification of a large number of samples with SWATH. In contrast, PAcIFIC data can be directly submitted for identification. Another advantage of PAcIFIC is the limited effect of co-fragmentation events attributable to narrow mass window isolation for precursor channels and a larger dynamic range of identification. This also facilitates the computation of precursor ions based on MS/MS fragmentation patterns. Generally, it is admitted that data acquisition in all data-independent strategies (SWATH, PAcIFIC, MS^E, etc.) is highly reproducible. These strategies are thus particularly relevant for quantitative proteomics. Throughout the manuscript, the term PAcIFIC designates a small (typically 3 *m/z* units) precursor mass window and multiple injections,

whereas DIA refers to the more general data independent strategy.

To maximize the outcome from DIA data, the development of dedicated software is necessary. DIA data volume can be large, mostly due to the redundancy of the MS/MS data. Many of the acquired spectra contain fragment ions of several co-eluting peptides at comparable MS/MS signal intensities. The proportion of chimeric MS/MS spectra was indeed estimated to 4 % for a typical proteomics experiment of medium complexity [15], whereas other estimates indicate higher levels of up to 20 % [16, 17]. Deconvolution of these chimeric spectra as well as clustering replicate spectra increase the accuracy of the database identification [17]. Another problem of peptide identification from DIA data is that the precursor ion m/z value is only approximately known, since it lies somewhere within the m/z window used to isolate this particular ion. In 2006, Venable and co-workers described a method to compute the m/z of precursor ions from MS/MS spectra in order to improve the precursor mass precision in low resolution instruments [18]. An alternative way of computing more accurate precursor ion m/z is to implement a full MS survey scan before each cycle of MS/MS spectra and use this information to assign precursors ion m/z values in the corresponding MS/MS spectrum [19]. But this approach is not always optimal with PAcIFIC data because precursor ions are often not visible in the survey scan [11].

Algorithms to compress and enhance DIA spectra have already been published [17], but these methods rely heavily on smooth and clear elution profiles of fragment ions. However, elution profiles are often noisy, especially for methods that use small incremental windows where only a few fragmentation spectra are measured during elution of a peptide. In contrast to existing methods, our approach makes use of both the time and m/z dimension to obtain a more accurate grouping of spectra. Such an approach is more versatile to the variation of precursor ion isolation window and overcomes the issue with imperfect elution profile of peptides. The first step consists in detecting local maxima from extracted fragment ion chromatograms. In the second step, spectra with elution times in the vicinity of a local maximum are clustered according to their pairwise similarity using an algorithm based on network clustering. This second step is crucial in order to form proper consensus spectra. We show that processing PAcIFIC MS/MS data with this algorithm increases the number of identifications and reduces the total number of MS/MS spectra and peaks submitted to the database search. In addition, we investigate the potential of an algorithm based on the complementarities of C- and N-terminal fragment ions to compute precursor ion m/z values from PAcIFIC MS/MS spectra.

Experimental

Materials

Iodoacetamide (IA) and acetonitrile were purchased from Sigma (St. Louis, MO, USA). Urea, ammonium bicarbonate

(AB), dithioerythritol (DTE), and water for chromatography and dilution were from Merck (Darmstadt, Germany). Porcine trypsin and formic acid (FA) were, respectively, from Promega (Madison, WI, USA) and Biosolve (Valkenswaard, The Netherlands). Stationary phases for columns were from Michrom (Auburn, CA, USA). Analytical column (o.d. = 375 μ m, i.d. = 75 μ m, L = 150 mm) and pre-column (o.d. = 375 μ m, i.d. = 100 μ m, L = 20 mm) was made from fused silica tubing from BGB Analytik AG (Boeckten, Switzerland). Ultrasonicator was from Hieschler Ultrasound Technology (Teltow, Germany).

Sample Preparation

Soluble proteins from MCF-7 cells were extracted by ultrasonication (Ultrasonic processor UIS250V; Teltow, Germany) and centrifugation at -4 °C. The supernatant was used for liquid digestion in 6 M UREA and 50 mM BA; 38 mM DTE was added and the solution was incubated at 37 °C for 60 min. Then, 108 mM IAA was added for alkylation during 60 min in the dark. Liquid digestion was performed overnight, by adding 1/50 ratio of proteins/trypsin. The digested solution was desalting with a C18 micro-spin column (Harvard Apparatus, Holliston, MA, USA) and dried. Dried material was suspended in CH3CN/FA 5 %/0.1 %.

Liquid Chromatography-Mass Spectrometry

The LC-MS/MS system consists of a NanoAcuity chromatograph (Waters, Milford, MA, USA) interfaced with an LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific, San Jose, CA, USA). Peptides were trapped on a homemade, 20 mm long precolumn of 100 μ m i.d. and separated on a 150 mm analytical column of 75 μ m inner diameter. The analytical separation was run for 65 min using a gradient of H₂O/FA 99.9 %/0.1 % (solvent A) and CH₃CN/FA 99.9 %/0.1 % (solvent B). The gradient was run as follows: 0–1 min 95 % A and 5 % B, then to 65 % A and 35 % B at 55 min, and 20 % A and 80 % B at 65 min at a flow rate of 220 nL/min. For PAcIFIC tandem mass spectrometry, full MS spectra were acquired in the Orbitrap detector from m/z =400–2000 before each cycle over precursor ion channels. The target ion population was 500,000 ions. MS/MS spectra were acquired over 20 precursor ion channels in the linear ion trap for each LC-MS/MS analysis, with an isolation window of \pm 1.5 m/z units, a channel's increment of 2.0 m/z units, NCE=35 % for CID and, target ion population of 10,000 ions. In total we covered a precursor m/z unit range of 430–1308 and injected the same sample 22 times. For example, the first injection (fraction 1) of our PAcIFIC data set covers precursor m/z from 430–468 with an overlap of 1 m/z units between the 20 consecutive precursor channels because of the applied isolation window for each precursor channel. The second fraction covers precursor m/z range from 470 to 508 units.

Data Analysis

As previously mentioned, one of the advantages of the PAcIFIC method is that one can directly submit the spectra for peptide identification. The spectra are processed independently for each channel and fraction and the results for searching processed spectra in a database (DB) can be directly compared with the results obtained with non-processed PAcIFIC data. The general workflow of our MS/MS processing strategy can be divided into three steps (see Figure 1): (1) binning of fragment ions and peak filtering (noise removal), (2) Local maxima extraction from aligned extracted fragment ion chromatograms (FICs), and (3) data clustering of spectra by their similarities, merging, and consensus spectra building. These steps are now described in detail.

Peak Filtering and Binning MS/MS Spectra

The next step consists of grouping MS/MS spectra per m/z channel and removing peaks that have no positive effect on database identification. First, we erase noisy peaks in MS/MS spectra by using a filter that slides a given m/z window (10 m/z width) over the entire m/z range and retains only

peaks within 1.5 m/z units of the four top abundant peaks within the window. This deletes small and noisy peaks from MS/MS spectra and improves data clustering and identification performance. MS/MS spectra are grouped according to their precursor channels m/z within the current fraction (e.g., $m/z 858 \pm 1.5 m/z$) (see Figure 1a, b). Then for each MS/MS spectrum of a channel peaks are extracted with intensities, retention times, and m/z values for binning (bin size of 0.06 m/z units) (see Figure 1c). For each bin an extracted FIC is calculated by averaging the values of all intensities of the peaks that fall within the m/z bin. To generate these extracted FICs only fragment ions between 130 and 1600 m/z are considered. A total of 24,500 bins of size 0.06 m/z are obtained and the same number of FICs per precursor channel (including empty bins). The binning step can be seen as the decomposition of the precursor channel total ion chromatogram (TIC) into extracted FICs.

Local Maxima Extraction

First, FICs are extracted for each fragment m/z bin. The algorithm detects local maxima in extracted FICs and saves

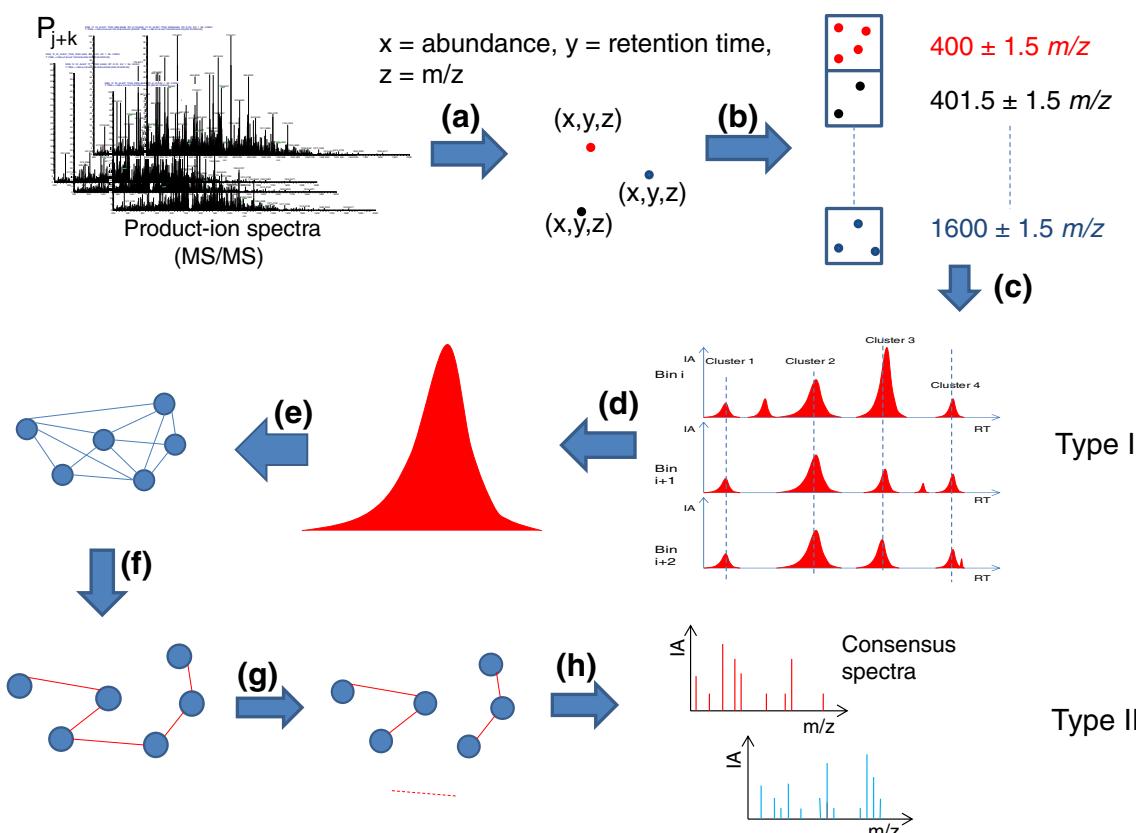


Figure 1. Tandem mass spectra are divided according to their precursor channels, then processed (a). From a set of spectra, peaks are extracted for binning (b) and extracted FICs are generated for selecting local maxima (c). An alignment of extracted FICs is performed to extract profiles according to retention time. All spectra in the vicinity of the center of a profile are taken (d) to generate a network of spectra based on their similarities (e). Minimum spanning tree algorithm is applied to find the shortest path (f) and the network is partitioned according to a cut-off (g). After the spectral network classification, linked spectra are merged together and consensus spectra are generated (h).

the coordinates of their apex (m/z , retention time and intensity) and corresponding bounds (left and right bounds for the elution time represent the start and end of the elution peak). Second, all local maxima that are close together ($\Delta t = \pm 3\text{--}4$ s) in the chromatographic separation are grouped. The left and right bounds of a group are set to the most frequent left and right bounds of all local maxima within this group. Finally for every group, a consensus spectrum is calculated by combining all spectra that elute within the respective bounds (consensus spectrum type I).

Clustering and Consensus Spectra

At the end of the local maxima extraction step, there are n groups of local maxima. In the absence of co-elution, all spectra within the bounds of a group should originate from the same analyte and, therefore, have similar peaks and relative intensities. However, to account for the possibility of co-elution of different analytes, the spectra are only merged after a further clustering step. All spectra within the bounds are compared with each other by means of a normalized dot product score (see Equation S-1) and the resulting score values are stored in a matrix. Then, a spectral network is built from this matrix (Figure 1e). This network is composed

of spectra (nodes) and similarity scores (edges). Prim's minimum spanning tree algorithm [20] is applied (Figure 1f) in order to find the tree, which links all nodes present in the network and where the total weights of edges is maximized (maximal total similarity). Then, the network is partitioned according to a similarity cut-off threshold (0.2) (Figure 1g), and linked spectra are merged to build consensus spectra (consensus spectra type II) (Figure 1h). Another advantage of this second step clustering is the possibility of defining the clustering granularity by a user-defined parameter.

Precursor Ion m/z Calculation

In DIA, the exact m/z of precursor ion is unknown because of the use of m/z channel isolation instead of individual isolation of precursor ions. We tried to solve this problem by implementing an approach based on the complementarities of N-terminal b-ions and C-terminal y-ions (Figure 2) rather than looking for a potential precursor ion peak in a full MS1 survey scan. This strategy also allows the processing of “orphan peptides” that are not detectable in MS1 scans. Assuming the precursor charge is two, the value of neutral precursor mass M_p can be calculated by summing the value of a singly charged y-ion and its complementary

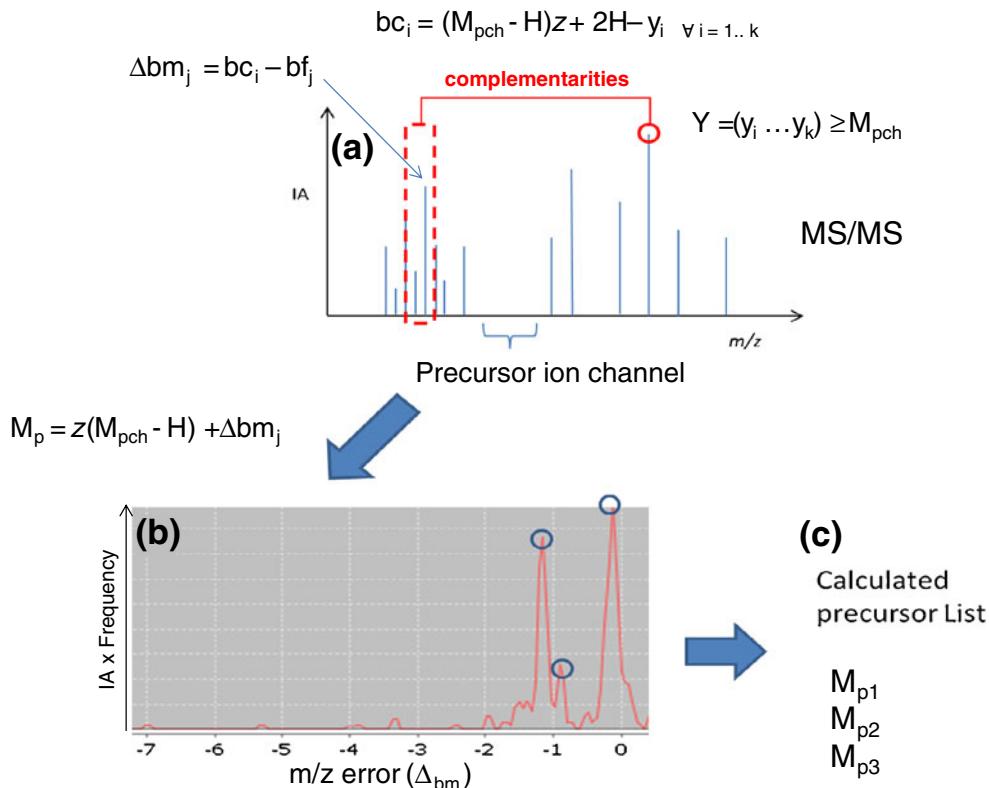


Figure 2. Overview of precursor ion correction. For a given spectrum, the precursor ion m/z value is estimated by the complementarities of b- and y-ions. **(a)** The center of precursor channel is used to calculate b-ion, and all possible shifts within a given window around this b-ion value are reported and binned to generate a plot based on the frequency of potential precursor ions found at a given shift **(b)**. The n most frequent precursor m/z values are then extracted **(c)**

singly charged b-ion (see Equation 1), where H is the mass of a proton.

$$M_p + 2H = b + y \quad (1)$$

This equation is adapted to PAcIFIC data as follows: for a given spectrum (in our case, a consensus spectrum) only a user-defined number of highest peaks are considered for the calculation. Then for all peaks ($Y = [y_1 \dots y_k]$), which have an m/z value greater than the m/z value of the precursor channel (M_{pch}), the algorithm tries to find the potential complementary b-ion ($bc_i \dots bc_k$) by subtracting the m/z value of y-ions from the value of M_p , which is roughly $M_{pch} - H$ multiplied by z , where z is the assumed precursor charge ranging here from 2 to 3.

$$bc_i = (M_{pch} - H)z + 2H - y_i \quad i = 1..k \quad (2)$$

The result of this operation gives the m/z values ($bc_i \dots bc_k$) where potential complementary b-ions can be found, assuming both b- and y-ions are singly charged. For all fragments that fall within a m/z window ($|bc_i - bf_j| \leq 1.25 m/z$) around the value bc_i , the m/z deviation values ($\Delta bm_1, \dots, \Delta bm_r$) between bc_i and the fragment m/z value (bf_1, \dots, bf_k) together with the intensity of the fragments are stored (see Figure 2a).

$$\Delta bm_j = bc_i - bf_j \quad j = 1 \dots r \quad (3)$$

Each Δbm_i is associated with a neutral precursor mass M_p via Equation 4.

$$M_p = z(M_{pch} - H) + \Delta bm_j \quad (4)$$

These operations are performed for all peaks with m/z values larger than M_{pch} present in a spectrum, and for every Δbm bin ($0.3 m/z$) the intensities of the peaks are summed. Finally, bins with the highest intensity values are chosen to calculate the possible precursor mass according to Equation 4 (Figure 2b, c). For more details, see the pseudo code in Supplementary Figure S-1.

Java and Dependent Libraries

All software were coded in Java. Most Java classes used to build these algorithms are available in Java Proteomics Library (JPL 1.0) developed at the SIB Swiss Institute of Bioinformatics. This library is freely available on [www.javaproplib.sourceforge.net](http://javaproplib.sourceforge.net). It contains classes and interfaces to facilitate the processing of data acquired in a mass spectrometer. It includes specific or generic parsers,

different types of filters for MS/MS spectra, similarity scoring systems, and more. Java Universal Network/Graph Frame (JUNG) is another library used to build our algorithm. It contains all classes to build and partition a graph. We used JUNG 2.0.1, which is available on <http://jung.sourceforge.net>.

Peptide and Protein Identification

Peak lists were generated from raw data using ReadW (<http://sourceforge.net/projects/sashimi/files/>). Peaklist files were searched against the UniProtKB/SwissProt database (2011_02 of 08-Feb-2011) using EasyProt (ref) [21, 22] (GeneBio, Geneva, Switzerland). *Homo sapiens* taxonomy was specified for database searching. The parent ion tolerance was set to 1.3 Da (this value gives highest number of identification on tested fractions) for PAcIFIC. Variable amino acid modifications were oxidized methionine and carbamidomethylated cysteine. Trypsin was selected as the enzyme, with one potential missed cleavage, and the normal cleavage mode was used. The peptide P value was 0.05 for LTQ. False discovery rates (FDR) were estimated using a reverse decoy database [23]. All datasets were searched once in the forward and once in the reverse database separately. Protein and peptide score thresholds were then set up to maintain the FDR below 5 %. For this analysis, only proteins matching two different peptide sequences were kept.

Results and Discussion

Data Reduction

To measure the effect of data reduction, we simply compared the number of spectra and peaks submitted for identification with and without data filtering and clustering (Figure 3). More than 18 % of MS/MS spectra and 55 % of peaks were removed after data processing. Peak number reduction is mainly due to peak filtering that cleans MS/MS spectra prior to binning. Binning also reduces the number of peaks but to a lesser extent. The reduction of the number of spectra submitted for identification takes place during the second step of local maxima extraction and type II consensus spectra building. This reduction depends only weakly on the parameters used to calculate spectral similarity (e.g., m/z tolerance for peaks alignment) and the cut-off value used after the MST algorithm (see Supplementary Figure S-2).

Identification

As described above, consensus spectra of type I are generated in the first step by considering only those peaks with similar local maxima (apex of an elution profile relative to ion chromatogram) and ignoring all others. Identification of the MS/MS spectra by EasyProt showed 7471 unique

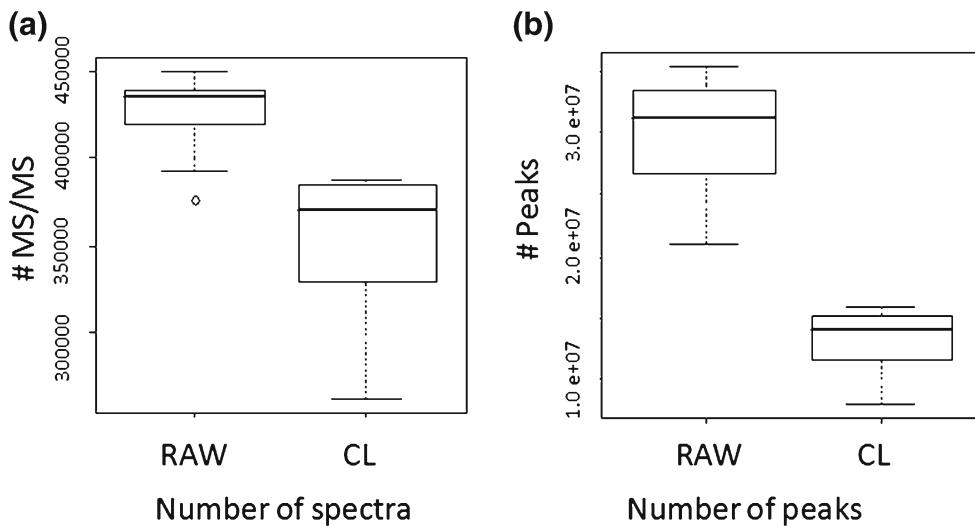


Figure 3. Data reduction after type II clustering (CL). The MS/MS spectra and peak counts are obtained from all 22 fractions. The boxes show the median, lower and upper quartiles of all 22 samples. Extreme values are indicated by horizontal lines and outliers by small circles. **(a)** The number of MS/MS spectra and **(b)** number of peaks submitted to data base search

peptides and 1157 protein identifications with at least two unique peptides and a FDR of 5 %. However, these values were lower than the results that we obtained with non-processed PAclFIC data, which showed 8411 unique peptide and 1247 protein identifications. Data inspection revealed that more than one peptide was frequently found in the vicinity of a local maximum. In such cases, the grouping of spectra based on local maxima is too coarse and consensus spectra of type I do not increase the number of identifications. Our approach can be improved if we take into account the potential co-elution of different analytes around a given local maximum. This procedure is similar to the one described by Frank and co-workers [24], but only applied to the subset of spectra in the vicinity of the local maximum. After spectral network partitioning of MS/MS spectra within the elution profile the resulting type II consensus spectra were submitted to database search for identification. We

identified 8925 unique peptides and 1399 proteins with at least two unique peptides and an FDR of 5 % (Figure 4). This corresponds to an increase of ~7 % of unique peptides and proteins relative to nonprocessed data. This increase is in agreement with the 4 %–5 % of chimeric spectra found by Scherl and co-workers [15] based on MS/MS identifications.

Type II consensus spectra led to an increase of 15 % and 19 % in unique peptide and protein identification compared with type I consensus spectra. This difference is explained by the coarse effect of step I data clustering. A better decomposition of MS/MS spectra is obtained by using step II data clustering. Supplementary Figure S-3 shows a total ion chromatogram (TIC) of precursor ion channel 822 m/z . Even though no clear peaks are visible in this TIC, identification results show that several peptides elute during this time. Type II processed MS/MS spectra matched six peptides (among them one is built from two spectra and

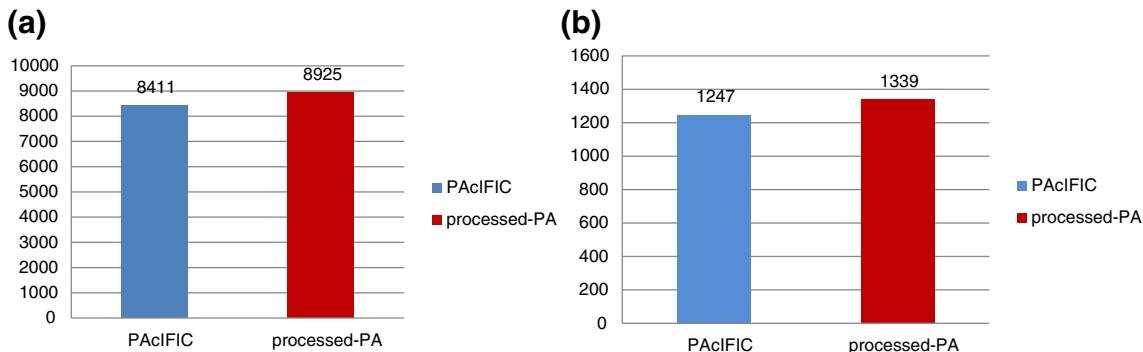


Figure 4. Unique peptide at a FDR of 5 % and protein identifications with at least two different peptide hits. Left **(a)**: unique peptide identifications. Right **(b)**: protein identifications. Processed PAclFIC (processed-PA) data with type II data clustering shows highest number of identifications for both peptides and proteins

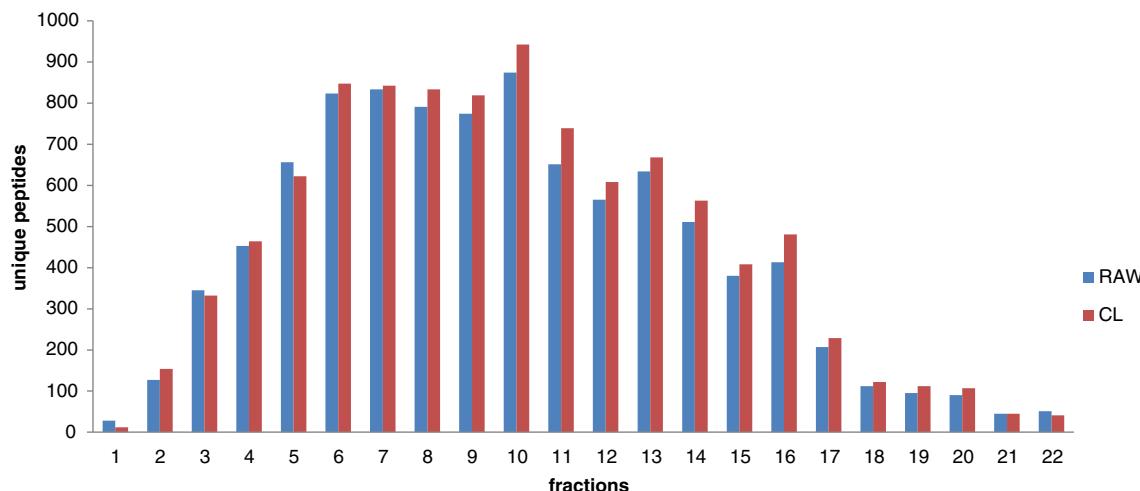


Figure 5. Distribution of the number of unique peptides identified in each fraction. The number of peptides is obtained at a false discovery rate <5 %. In most cases processed data show more identifications

specific to type II) whereas nonprocessed MS/MS spectra matched only two peptides. The higher number of identifications is due to the better quality of the spectra after peak filtering and type II clustering. The number of unique peptide identifications per fraction is displayed in Figure 5. A gain is observed over 22 fractions, even if some fractions in the low (fractions 1, 3, 4) and high *m/z* (fraction 22) regions display a slight decrease or equal number of peptides for processed and unprocessed data. One can notice the gain for the middle part of the fractions (relative to low and high

m/z region) where most of the tryptic peptides are observed with optimized CID activation energy. As mentioned by Scherl and co-workers for gas phase fractionation, the low and high *m/z* regions were not optimized for conventional CID activation in shotgun proteomics. Further, we counted the number of spectra that yielded exactly 1, 2 or 3 unique peptide matches per spectrum. The general trend shows that type II clustered spectra produced more single peptide hits (Table S-1, Supplementary Figure S-4A) and less hits to multiple peptides (Supplementary Figures S-4B, C). As

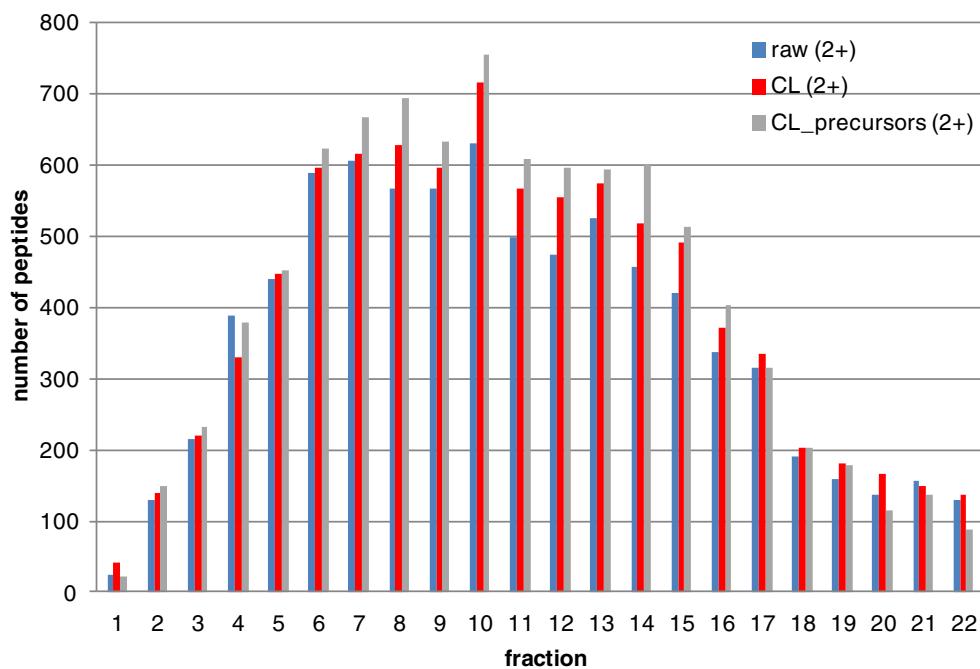


Figure 6. Distribution of unique peptides per fraction for doubly charged spectra. In blue, results from raw MS/MS spectra (raw [2+]). In red, results from MS/MS spectra after data clustering (type II) (CL [2+]). In grey, results after data clustering and precursor correction (CL_precursor [2+]). In the majority of fractions, CL_precursor [2+] display higher number of peptides identification rates

previously mentioned, the data clustering seems to work less efficiently for the higher mass region. This can be a matter of clustering parameters that need to be adapted for these regions. These additional data support the ability of our algorithm to correctly cluster and decompose MS/MS spectra from PAcIFIC data.

Precursor Ion m/z Calculation

One of the most common problems in DIA is that the exact precursor ion m/z ratio and charge state are often unknown. Panchaud and co-workers reported that at least 30 % of identified peptides with PAcIFIC have nonidentifiable precursor ions in the survey scan (called “orphan peptides”). Some groups worked at detecting precursor ions in the MS1 survey scan and assigning the detected m/z value to the corresponding MS/MS spectra. The latter strategy is limited by the intraspectrum dynamic range of the mass spectrometer and offers no solution for orphan peptides. In order to increase the accuracy of identified precursor ions m/z even for orphan peptides and to reduce the time spent for database search, we developed an algorithm that uses the complementarities of N- and C-terminal fragment ions to calculate the m/z of precursor ions. In the absence of information about the precursor ion charge state, each MS/MS spectrum is duplicated and searched for doubly and triply charged precursor m/z values against the database. In this preliminary test, we only considered doubly charged precursors and selection of the two most intense precursor ion peaks. Figure 6 displays the distribution of identified doubly charged peptides per fraction with FDR=5 %. The data combined with data clustering and precursor ion correction (CL_precursor_[2+]) shows most unique peptide identifications (an increase of 12 % compared with data without data clustering and precursor ion correction). The improvement of precursor ion m/z accuracy can be observed in Supplementary Figure S-5. The deviation between predicted and theoretical m/z values clearly becomes more concentrated around smaller values.

Conclusion

We presented a method to process DIA data that allows increasing the number of peptide and protein identifications while decreasing the data size. For PAcIFIC data this approach showed an increase of 7 % for both peptide and protein identifications. The number of submitted MS/MS spectra was reduced by 18 % and 55 % of the peaks were discarded. In addition, we attempted to compute the precursor ion m/z from MS/MS spectra and produce a comprehensive method to process DIA data. The results show an improvement of precursor ion m/z accuracy and a gain of unique peptide identifications for all PAcIFIC fractions after applying a mass corrective function. The precursor mass correction did not work for all spectra but we believe that there is room for improvement and anticipate a

better version of the algorithm. The spectrum clustering algorithm also determines when the peptides elute, which is of immediate importance for peptide quantification based on DIA data. In future studies, we would like to explore the potential of the DIA data processing pipeline for quantitative proteomics.

Acknowledgment

The authors thank the Swiss National Science Foundation (SNSF), grant 315230_130830, for support of this work. The authors declare no conflict of interest.

References

- Gatlin, C.L., Eng, J.K., Cross, S.T., Detter, J.C., Yates, J.R.: Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Anal. Chem.* **72**, 757–763 (2000)
- Washburn, M.P., Wolters, D., Yates III, J.R.: Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001)
- Chang, E.J., Archambault, V., McLachlin, D.T., Krutchinsky, A.N., Chait, B.T.: Analysis of protein phosphorylation by hypothesis-driven multiple-stage mass spectrometry. *Anal. Chem.* **76**, 4472–4483 (2004)
- Liu, H., Sadygov, R.G., Yates III, J.R.: A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004)
- Purvine, S., Eppel, J.-T., Yi, E.C., Goodlett, D.R.: Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* **3**, 847–850 (2003)
- Silva, J.C., Gorenstein, M.V., Li, G.-Z., Vissers, J.P.C., Geromanos, S.J.: Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteom.* **5**, 144–156 (2006)
- Venable, J.D., Dong, M.-Q., Wohlschlegel, J., Dillin, A., Yates, J.R.: Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1**, 39–45 (2004)
- Panchaud, A., Scherl, A., Shaffer, S.A., von Haller, P.D., Kulasekara, H.D., Miller, S.I., Goodlett, D.R.: PAcIFIC: how to dive deeper into the proteomics ocean. *Anal. Chem.* **81**, 6481–6488 (2009)
- Yi, E.C., Marelli, M., Lee, H., Purvine, S.O., Aebersold, R., Aitchison, J.D., Goodlett, D.R.: Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis* **23**, 3205–3216 (2002)
- Spahr, C.S., Davis, M.T., McGinley, M.D., Robinson, J.H., Bures, E.J., Beierle, J., Mort, J., Courchesne, P.L., Chen, K., Wahl, R.C., Yu, W., Luethy, R., Patterson, S.D.: Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. I. Profiling an unfractionated tryptic digest. *Proteomics* **1**, 93–107 (2001)
- Panchaud, A., Jung, S., Shaffer, S.A., Aitchison, J.D., Goodlett, D.R.: Faster, quantitative, and accurate precursor acquisition independent from ion count. *Anal. Chem.* **83**, 2250–2257 (2011)
- Chen, S., Panchaud, A., Goodlett, D., Shaffer, S.: Making a case for data-independent tandem mass spectrometry workflows. *J. Biomol. Tech.* **21**, S52–S53 (2010)
- Hengel, S.M., Murray, E., Langdon, S., Hayward, L., O’Donoghue, J., Panchaud, A., Hupp, T., Goodlett, D.R.: Data-independent proteomic screen identifies novel tamoxifen agonist that mediates drug resistance. *J. Proteome Res.* **10**, 4567–4578 (2011)
- Gillet, L.C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., Aebersold, R.: Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteom.* **11**, (2012)
- Scherl, A., Tsai, Y.S., Shaffer, S.A., Goodlett, D.R.: Increasing information from shotgun proteomic data by accounting for misassigned precursor ion masses. *Proteomics* **8**, 2791–2797 (2008)
- Ahrné, E., Ohta, Y., Nikitin, F., Scherl, A., Lisacek, F., Müller, M.: An improved method for the construction of decoy peptide MS/MS spectra suitable for the accurate estimation of false discovery rates. *Proteomics* **11**, 4085–4095 (2011)

17. Bern, M., Finney, G., Hoopmann, M.R., Merrihew, G., Toth, M.J., MacCoss, M.J.: Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Anal. Chem.* **82**, 833 (2010)
18. Venable, J.D., Xu, T., Cociorva, D., Yates III, J.R.: Cross-correlation algorithm for calculation of peptide molecular weight from tandem mass spectra. *Anal. Chem.* **78**, 1921–1929 (2006)
19. Carvalho, P.C., Han, X., Xu, T., Cociorva, D.: da G. Carvalho M., Barbosa, V.C., Yates, J.R., 3rd: XDIA: improving on the label-free data-independent analysis. *Bioinformatics* **26**, 847–848 (2010)
20. Prim, R.: Shortest connection networks and some generalizations. *Bell Syst. Technical J.* **36**, 1389–1401 (1957)
21. Gluck, F., Hoogland, C., Antinori, P., Robin, X., Nikitin, F., Zufferey, A., et al.: EasyProt—an easy-to-use graphical platform for proteomics data analysis. *J. Proteom.* **79**, 146–160 (2013)
22. Colinge, J., Masselot, A., Giron, M., Dessingy, T., Magnin, J.: OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* **3**, 1454–1463 (2003)
23. Elias, J.E., Gygi, S.P.: Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007)
24. Frank, A.M., Bandeira, N., Shen, Z., Tanner, S., Brigg, S.P., Smith, R.D., Pevzner, P.A.: Clustering millions of tandem mass spectra. *J. Proteome Res.* **7**, 113–122 (2008)