Data-independent mass spectrometry for the analysis of peptide mixture

Pak, Hui Song

UNIVERSITE DE GENEVE

Section des sciences pharmaceutiques

FACULTE DES SCIENCES

Prof. Denis F. Hochstrasser

Département des sciences des protéines humaines

FACULTE DE MEDECINE

Dr Alexander Scherl

Data-independent mass spectrometry for the analysis of peptide mixture

THESE

Présentée à la faculté des sciences de l'université de Genève pour obtenir le grade de docteur ès sciences, mention interdisciplinaire

par

HuiSong Pak

de

Nyon (VAUD)

Thèse n° 4764

Genève

2015

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES SCIENCES

Doctorat ès sciences
Mention interdisciplinaire

Thèse de *Monsieur Hui Song PAK*

intitulée :

"Data-independent Mass Spectrometry for the Analysis of Peptide Mixture"

La Faculté des sciences, sur le préavis de Monsieur D. HOCHSTRASSER, professeur ordinaire et directeur de thèse (Section des sciences pharmaceutiques et Faculté de médecine, Département de biologie structurale et bioinformatique), Monsieur A. SCHERL, docteur et codirecteur de thèse (Faculté de médecine, Département de biologie structurale et bioinformatique), Monsieur Y. KALIA, professeur associé (Section des sciences pharmaceutiques), Monsieur M. Müller, docteur (Département d'informatique, Institut Suisse de Bioinformatique) et Monsieur C. MASSELON, docteur (Etude de la Dynamique des Protéomes, Institut de Recherches en Technologies et Sciences pour le Vivant, Commissariat à l'énergie atomique et aux énergies alternatives, Grenoble, France), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 19 janvier 2015

Thèse - 4764 -

Le Décanat

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

Résumé

La spectrométrie de masse en tandem couplé à la chromatographie liquide (LC-MS/MS) est largement utilisée pour identifier et quantifier les protéines. La méthode d'acquisition couramment utilisée est "l'échantillonnage dépendant des données" qui ne sélectionne que les peptides les plus abondants dans un échantillon. Pour faire face à cette sélection biaisée, une méthode alternative a été développée, dite "l'échantillonnage indépendant des données". Cette méthode, qui est de plus en plus utilisé, consiste à balayer systématiquement et de manière incrémentale une fenêtre de masse restreinte pour couvrir une gamme de masse donnée. Dans le cadre de ce travail, cette dernière méthode a été modifiée pour la quantification dirigée de certaines protéines. En définissant les fenêtres de masses à balayer systématiquement dans un spectromètre de masse de type piège à ion, les résultats ont montré que l'identification et la quantification de protéines sont possibles sur cinq ordres de grandeur dans un échantillon complexe. Cela a ainsi démontré que la sélectivité des ions précurseurs est augmentée, en particulier, si on utilise une fenêtre d'isolation restreinte pour ces derniers. Nous avons donc utilisé la méthode d'échantillonnage indépendant des données pour analyser le protéome des cellules humaines MCF-7 en culture. Pour augmenter le taux d'identification de spectres, nous avons développé un ensemble d'algorithmes pour traiter les spectres de masse issus de co-elution et de co-isolation de peptides de masse similaire, appelés spectres chimériques, et inhérent à cette méthode. L'application de ces algorithmes sur les données de spectrométrie de masse acquises en mode d'échantillonnage indépendant des données a montré une augmentation du nombre d'identification de protéine et une réduction de la taille des données. De plus, le nombre de spectres chimériques, *i.e.* contenant des ions fragments de multiple ions peptidiques, a été réduit démontrant ainsi que l'algorithme est capable de séparer ces fragments sur la base de leur profile d'élution chromatographique. Ensuite, cette approche combinée est utilisée pour étudier les modifications chimiques des protéines lors du

vieillissement des globules rouges dans des poches de transfusion sanguines. En effet, la qualité des globules rouges ainsi que leurs conservations sont importantes dans la médecine de transfusion afin de garantir une meilleure sécurité des patients, améliorer le tri des poches de sang et faire face à des pénuries chroniques. Afin d'obtenir un maximum d'informations, nous avons combiné les acquisitions d'échantillonnage dépendant des données et d'échantillonnage indépendant des données, ainsi qu'une recherche de peptides avec des modifications non-prévisibles. Une liste exhaustive de peptides, leurs quantités relatives, ainsi que les modifications peptidiques ont été rapportées. Ces données ont finalement été croisées afin de retenir uniquement les peptides contenant des modifications chimiques dont l'abondance varie fortement avec l'âge des globules rouges. Ces peptides représentent des marqueurs potentiels du vieillissement des globules rouges.

Abstract

Tandem mass spectrometry hyphenated to liquid chromatography (LC-MS/MS) is widely used for protein identification and quantification in proteomics. The commonly used ion sampling method is data-dependent acquisition (DDA), which consists in the selection of the most abundant peptides in a sample for tandem mass spectrometry. To overcome this biased selection of precursor ions, data-independent acquisition can be used. This ion sampling method consists in a systematic interrogation of all precursor ions within a fixed isolation window width and predefined list of isolation window increments to cover the desired mass range. For this thesis, the data-independent acquisition method was modified for targeted peptide quantifications using a systematic scanning of a predefined list of isolation window in an ion trap. The results showed a dynamic range of identification and quantification of five orders of magnitude in complex samples. It showed an increase of ion selectivity and ion sampling specificity with data-independent acquisition, especially when using narrow isolation windows. Considering these results, data-independent acquisition was used to analyze human MCF-7 cells in culture. The nature of the data-independent ion sampling method increases the observation of peptide co-elution with similar precursor ion *m/z*. This result with tandem mass spectra containing fragment ions from more than one precursor ion, also called chimeric spectra. A suit of algorithm was thus developed to process data-independent tandem mass spectra and to separate chimeric spectra. Application of this algorithm resulted with an increase of protein identifications and a reduction of tandem mass data. In addition, the number of chimeric tandem mass spectra was reduced, demonstrating the ability of the algorithm to separate fragment ions based on their chromatographic elution profiles. This combined approach was then applied to study chemical modification of proteins during red blood cell aging in blood storage bags. Indeed, the preservation and quality of red blood cells is important for transfused patients and to optimize the provision of blood stocks

in hospitals. In order to obtain proteome information as comprehensive as possible, the samples were analyzed in data-dependent and data-independent mode. After, an open modification search was performed to identify modified peptides. An exhaustive list of peptides, peptide expression profiles and modifications was reported. These data were integrated to retain only peptides containing chemical modification whose abundance varies with the age of red blood cells. These modified peptides represent potential biochemical indicators of red blood cell aging.

Remerciements

J'aimerais exprimer toute ma gratitude à ceux qui ont, d'une façon ou d'une autre, contribué à la réalisation de ce travail.

Aux membres du jury, Dr Christophe Masselon, Dr Markus Müller et Prof. Yogeshvar N. Kalia pour avoir accepté de lire et d'évaluer mon travail de thèse.

Aux Prof. Denis Hochstrasser et Dr Alexander Scherl, directeur et co-directeur de thèse, qui m'ont permis de travailler et d'évoluer au sein de leur équipe. Merci à Alex de m'avoir fait confiance et donné l'opportunité de réaliser ce travail. Sa vivacité d'esprit, son enthousiasme et sa patience m'ont aidé à avancer et m'améliorer tout au long de ces quatre années de thèse.

Au Dr Markus Müller, pour avoir supervisé la partie bioinformatique de ce travail et ses nombreux conseils pertinents. Sa bonne humeur et son ingénuité à toujours trouver une solution m'ont toujours agréablement surpris.

A tous les membres du TBG et CPC pour leur soutien, plus particulièrement aux Prof. Jean-Charles Sanchez et Dr Pierre Lescuyer.

A mes collègues doctorants Didia, Linnea, Leire, Francesco et collègues post-doc: Domitille, Floriane, Virginie, Affif, David et ainsi que Carla, Fabienne, Paola, Vanessa, Alex et Julien. Merci à vous pour les discussions scientifiques et les bons moments passés ensemble.

Au Triumvirat, pour les moments de fou rire et autres discussions autour d'un café. Parfois, il me semblait que le café avait un arrière-goût de houblon!

A mes amis, Dom, turtle et Co. pour les bons comme les mauvais moments

A mes parents, mon frère et ma sœur pour leur soutien. Merci à Jason, pour m'avoir remis sur le droit chemin.

A Tobby, le plus fidèle de nous tous, à qui je dédie cette thèse en souvenir de sa mémoire.

TABLE OF CONTENTS

1 Introduction

Protein identifications and quantifications are important to study the dynamic of a proteome [1]. This is the beginning for further investigations such as protein interactions in complex biological systems [2]: How these interactions are related to the function of a system and how the system behaves in the presence of such interactions. Consequently, protein identification and quantification is essential in systems biology to understand the complexity of living organisms. In the early days of proteomics, available techniques to sequence and identify proteins were limited to one or two-dimensional gel electrophoresis separation [2], Edman degradation sequencing [3] and western blotting [4][5]. All mentioned techniques were time consuming. Fast, high throughput, and on-line combination of separation technique were thus necessary. Especially, a request for an analytical system that allows a systematic and automated analysis of biological samples was demanded. The solution was the combination of a high performance separation technique such as high performance liquid chromatography (HPLC) and mass spectrometry (MS). Liquid chromatography allowed decreasing sample complexity before the injection of eluting compounds into the mass spectrometer for ion analysis [6][7]. This coupling was possible with the improvement of soft ionization technique such as electrospray ionization [8] (ESI) and the development of adapted HPLC plumbing systems with reduced flow rate (nano-LC).

Nowadays, the hyphenation of liquid chromatography (LC) with mass spectrometry, especially with two stages of mass spectrometry (tandem mass spectrometry) is widely used to identify and quantify proteins and/or peptides in complex mixtures [9][10]. Gain in sensitivity was possible by the combination of miniaturized ESI [11][12] devices and liquid chromatography columns with small internal diameter. Such developments improved the dynamic range (defined as the ratio between the most concentrated and most diluted identified compound quantity) of detected proteins in a complex mixture. In parallel, development of mass spectrometry sampling methods based on the automatic selection of the most abundant

precursor ions for tandem mass spectrometry, a process known as “data dependent acquisition” (DDA) made characterization of complex protein mixtures possible. However, DDA tandem mass spectrometry has several inconveniences. A lack of reproducibility and a bias towards abundant ions are reported for precursor ion selection. To overcome these issues, research for a more selective and sensitive method of data acquisition methods were developed: The result is data independent acquisition (Shotgun CID, MS^E, PAcIFIC) [13][14][15], selected reaction monitoring (SRM) [16] or pseudo-selected reaction monitoring (p-MRM) [17][18]. All these techniques will be detailed in the next paragraphs.

1.1 High performance liquid chromatography

High performance liquid chromatography (HPLC) is a separation technique based on the chemical properties of macromolecules in a liquid sample. There are different methods of separation and most of them are based on affinity, polarity and size of the molecules. The affinities between mobile phase, stationary phase and the analyte are the main interaction involved for a HPLC experiment. Once trapped, the compound is eluted by a mobile phase with more affinity to the trapped compound compared to the stationary phase. The elution can be done in isocratic or gradient mode. HPLC became popular because it was simple and offered several interaction types to isolate components of interest from a complex mixture.

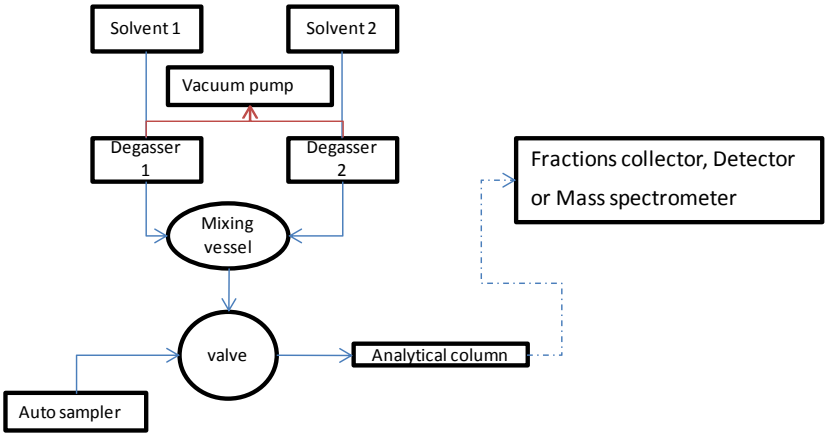


Figure 1: Schematic representation of a HPLC system. Important pieces that compose HPLC are shown. A vacuum pump serves to operate as solvent degasser.

The system is composed of 1) an auto sampler to inject samples, 2) a pump that pulses solvents (mobile phase) into the system, 3) columns (stationary phase) to trap analytes according to their interactions and 4) a detector, for example a UV detector or a mass spectrometer. A HPLC system functions as following: First, the sample is injected into the sample loop via an autosampler (not shown in **Figure 1**). Then the pump pulses the mobile phase (solvents) through the loop and the sample is pushed through the columns. Depending on the type of interactions, different mobile phases can be used to bind and elute analytes from the mobile phase to the stationary phase. The detector monitors retention time and relative quantity of a given compound.

The chromatographic separation is visually characterized by response-signals over time. The quality and separation depend respectively on the size of the resin particles (stationary phase) and the length of the column for a given stationary and mobile phase. The height of theoretical plates (H) measures the efficiency of a column and is determined by the number of theoretical plates (N) and the length of the column (L):

$$\frac{L}{N} = H \text{ [19]}$$

If N is not known, it is possible to calculate it by using the width at half height $W_{1/2}$:

$$N = 5.54 \times \frac{T_R^2}{W_{1/2}^2}$$, with T_R as retention time [19]

Higher N values signify better separation efficiency. This means also that smaller H values provides better efficiency. Another important parameter is the resolution (R) of an HPLC. The resolution is the ability of the system to separate two peaks. The corresponding formula is:

$$R = 2 \times \frac{T_{R2} - T_{R1}}{W_1 + W_2}$$, with W as the width at baseline [19]

Usually the measure of W is done at 10-13% of the peak intensity [19]. It is considered that two peaks are resolved when the R value is larger than 1.5 or 2. The peak intensity and the area of the peak reflect the concentration or the amount of injected sample. This relationship is only valid under the hypothesis of Gaussian peak shape intensity.

1.1.1 Reverse phase liquid chromatography

Historically, HPLC was performed with a polar and hydrophilic stationary phase, silica matrix with OH free groups (Si-OH), able to form hydrogen bonds with analytes. In contrast, the term reverse phase is used to describe a stationary phase less polar than the mobile phase [19]. Such phases can be obtained with silanol groups of silica covalently bound to C18 or C8 alkyl chains. The mobile phase is usually composed of a mixture of water with various water-miscible solvents (ex: methanol, acetonitril and isopropanol). The miscibility of such solvents in water offers the ability to use various gradients and elution profiles. As gradient elution is often involved in reverse-phase chromatography, the solvents should be as pure as possible, in order to minimize trace impurities to be adsorbed and eluted (giving rise of ghost peaks formation in the chromatogram). Usually, with reverse-phase chromatography methods, hydrophilic or polar compounds are eluted first. Then, by decreasing the polarity of the

mobile phase, hydrophobic analytes are eluted. The reverse order of compound elution compared to normal phase is observed. The advantage of RP-HPLC in bioanalytical sciences is obvious. Samples are frequently in aqueous solution, they often have a long backbone chain of carbons and are thus frequently more hydrophobic than water and well suited for RP-HPLC . The research to improve columns separation performances led to different and more efficient strategies of protein/peptide detection. One of the ways to improve the sensitivity was the miniaturization of columns. The formula of factor's sensitivity F~ $\frac{d_1^2}{d_2^2}$ [20] showed that a downscaling from an internal diameter of 4.6 mm to 50 µm brought a gain in sensitivity of 8500. The gain in sensitivity is evaluated by comparing peak intensities. The formula was valid if other parameters remained constant [20]. With the miniaturization, polyphasic columns were tested, for example the combination of strong cation exchange (SCX) [21] and reverse phase (RP) [22] to bind and elute sequentially hydrophobic proteins/peptides (2D-LC) [23]. The problem with such a strategy is the long analysis time required for proteins/peptides separation on more than one stationary phase. In order to decrease analysis time, a plumbing system called vented columns was invented (see **Figure 2**). The vented column was originally designed for liquid chromatography (LC) coupled with electrospray ionization mass spectrometry (ESI-MS). The plumbing geometry was popularized by Yates and co-workers in 1998 [24] but was already described in 1994 [25]. The system uses a short trapping column for fast binding of analytes with a relatively high flow rate. Analytes are then sequentially eluted from a trapping column to a longer column, the analytical column. The particularity of a vented column system is the presence of a vent (split) between the trapping and analytical column, in order to control the back-pressure and to increase the flow rate during sample loading. A sintered frit at the end of the pre-column and a tapered tip at the end of the analytical column prevent the loss of the stationary phase.

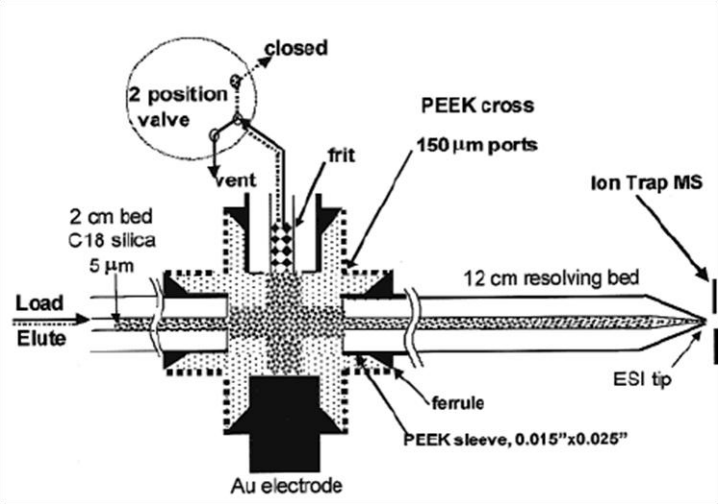


Figure 2: Vented column configuration. The flow rate is controlled by a switching valve system. Stationary phase of columns are the same and a sintered frit is built to retain trapping column's stationary phase. A gold electrode is fixed and an electric potential is applied on it to generate an electrospray for direct interfacing with a mass spectrometer. Reprinted from L. J. Licklider *et al.* [26].

The frit at the end of the pre-column is usually built by sintering underivatized silica beads [27] or by polymerization of KaSil and foramide [20]. In 2003, Yi and co-workers described a vented column (see **Figure 3**) which offered a higher flexibility *via* a trapping column packed with a different stationary phase as the analytical column [27]. It offered the ability to build independent trapping and analytical columns. Such configuration allowed analyzing more than 60 samples by LC-MS without losing sensitivity and without column replacement.

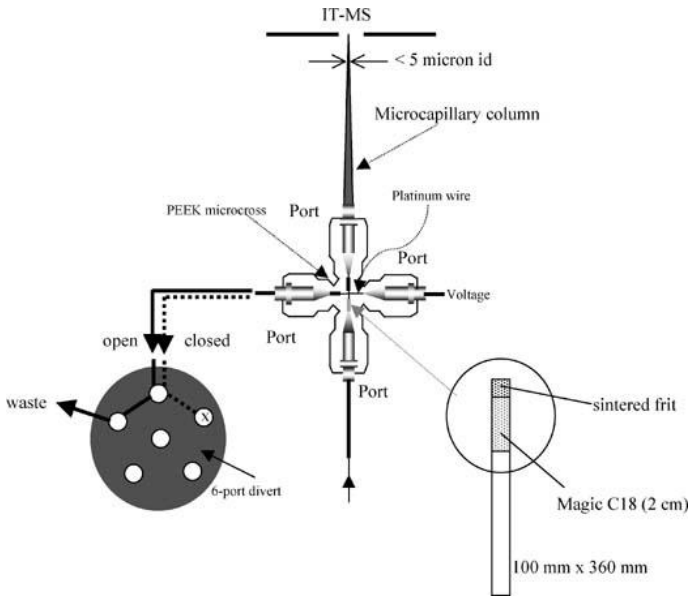


Figure 3: Schema of vented column with valve switching operation and a representation of a fritted trapping column. Reprinted from E. C. Yi, H. Lee *et al.*, [27].

1.2 Mass spectrometry

A mass spectrometer is an ion detector. The main function of mass spectrometry (MS) is to identify and quantify ionized atoms or molecules in the gas phase. Particularly a mass spectrometer measures the abundance of gas-phase ions by their mass-to-charge ratio (*m/z*). It generates a mass spectrum where the x and y axes correspond respectively to *m/z* and ion abundance. Typically a mass spectrometer is formed by 1) a sample introduction device, 2) a source to produce ions, 3) one or several mass analyzers, 4) a detector to measure the abundance of ions, and 5) a computer for data processing (see **Figure 4**).

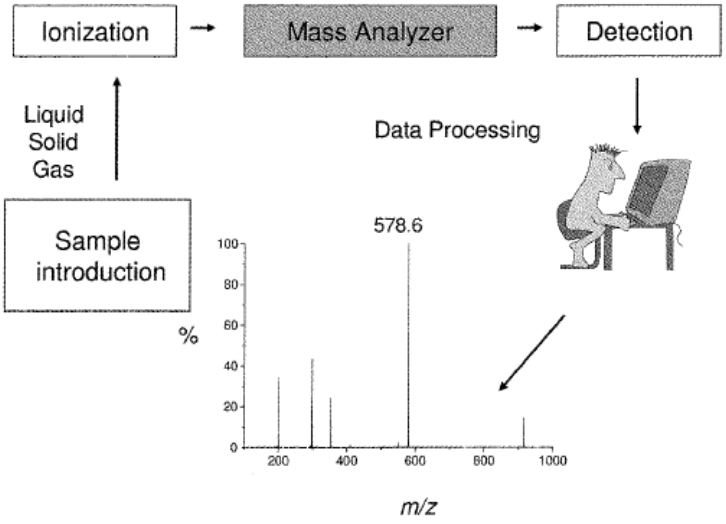


Figure 4: Schematical representation of the different devices involved in a mass spectrometer. The outcome of an analysis is shown as a spectrum. Reprinted from G. Hopftgarner *et al.*, [28].

Mass spectrometers usually operate in high vacuum in order to not disturb the trajectory of ions until they reach the detector. They can be grouped into 3 different operation modes: continuous (magnetic sector, quadrupole), pulsed (time of flight) and ion trapping devices (ion trap, Fourier transform ion cyclotron resonance and orbitrap) [28].

From an analytical point of view, some elements of MS analysis should be defined. This includes mass resolution, mass precision and mass accuracy. The mass resolution is given by

$$R = \frac{\Delta m/z}{m/z} \text{ [29]}$$

It means that two peaks of the same amplitude are resolved (distinct signals) if the valley between the two peaks is below 10% of the considered peak intensity. In practice, it is difficult to observe two peaks with exactly the same amplitude. So often the resolution is measured in one peak by the full width at half maximum (FWHM) and/or at FW 10% of the intensity:

$$R = \frac{\Delta m/z(\text{FWHM})}{\Delta m/z(FW\ 10\%)}$$ [29]

Mass accuracy is defined as the difference between theoretical *m/z* and measured *m/z* on instrument, and the accuracy of the instrument. The mass accuracy is often described in a relative manner, for example parts per million (ppm). Finally the mass precision is defined as root-mean-square deviation of a large number of repeated measurements [30][31]. **Figure 5** illustrates the difference between mass accuracy and mass precision. Another characteristic is the duty cycle of the mass spectrometer, defined as the part of ions of a particular *m/z* produced in the sources that are effectively analyzed. The unit can be expressed as a ratio or as a percentage. For instance 100% of duty cycle is reached with selected ion monitoring scan mode, whereas in simple scan mode (MS^1), the duty cycle decreases with the time spent to scan each ion [29].

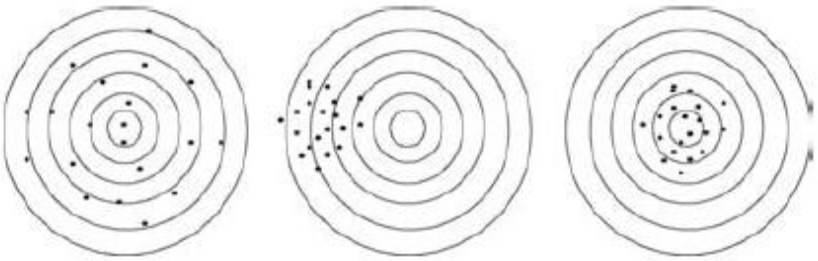


Figure 5: Left: low precision, but fair accuracy as the points are well distributed around the target. Centre: better precision but bad accuracy. Right: same precision as centre, but much better accuracy. Reprinted from E. de Hoffmann *et al.,* [29].

1.2.1 Ionization technique

To be analyzed by MS, an analyte needs first to be ionized into a stable gas phase ion. A classical ionization technique is electron impact ionization (EI). A heated filament that produces electrons (e^-) is in contact with vaporized molecules in a vacuum system (or at low pressure). The e^- are then accelerated to a given energy (eV, ionization energy). Gas-phase molecules are bombarded with e^-, and e^- are removed from molecules and radical cations

($M^{+\cdot}$) or multiply charged molecules are formed [32]. This ionization can be applied on solid, liquid or gas phase analytes as long as the molecules are stable during the vaporisation or sublimation process. Consequently, EI is limited to thermally stable and volatile molecules. Similarly, chemical ionization (CI) produces protonated or deprotonated molecules using the same ion source design than EI but with the presence of a reagent gas such as methane [33]. As in EI, the methane will react with an e^- and results into the formation of a radical cation $CH_4^{+\cdot}$. Then, the radical cation will react with other methane molecules, forming $CH_5^+ + CH_3^\cdot$. The ion CH_5^+ will react with the molecule or analyte (M) in acid-base type reaction: $M + CH_5^+ \rightarrow MH^+ + CH_4$. CI or EI is usually used for gas phase chromatography coupled with mass spectrometry (GPC-MS). Often, molecules must be derivatized first in order to be ionized and analyzed by mass spectrometry. Derivatization occurs when specific or unspecific molecules are modified by addition of chemical adducts in order to be adapted to a particular analysis (for example to make molecules more volatiles to be analyzed by GPC-MS).

1.2.2 Electrospray

Previously described ionizations methods were limited to small or thermally stable molecules, usually subjected to derivatization prior to GPC-MS. With the development of soft ionization techniques such as electrospray ionization (ESI) [8] or matrix assisted laser desorption ionization (MALDI) [34][35] and the ability to perform atmospheric pressure ionization (API) [36][37], it was possible to directly interface a classical HPLC system with MS. Soft ionization allows to ionize macromolecules such as proteins and peptides without extensive fragmentation. Among all ionization techniques, ESI is probably the most used one. This is due to its simplicity and its easy coupling to liquid chromatography (LC). Nowadays ESI is mostly performed at atmospheric pressure. A spray can be obtained with the use of 1) nebulising gas, 2) heating, 3) application of ultrasound, and 4) application of an electric field.

An electrospray is obtained by a potential applied between the ESI source and the counter electrode, and nebulized by a nebulising gas or high temperature. It is usually admitted that after the nebulisation, the charged droplets reduce their sizes by dividing (coulomb explosions), until ions are ejected from very small droplets as gas phase ions (see **Figure 6**). The evaporation of droplets into gas phase ions is not fully understood and is still under investigation. Commonly, two theoretical models are proposed: 1) the Ion Evaporation (IEV) model and 2) the Charge Residue (CR) model. It is reported that the IEV model explains better the evaporation of small molecules while the CR model is more favorable to explain proteins or peptide ionization [38]. The important feature of ESI is its sensitivity to concentration and not to the total amount of sample injected in the source [39]. This means that signal intensity is increased with decreasing flow rate. Exploiting this concentration dependence, micro-electrospray (µESI) or nano-electrospray (nESI) ionization which uses a flow rate of dozen nl min^{-1} was developed with adapted probe tips [25]. Another advantage of reducing the flow rate is less sample consumption and less contaminant from solvents.

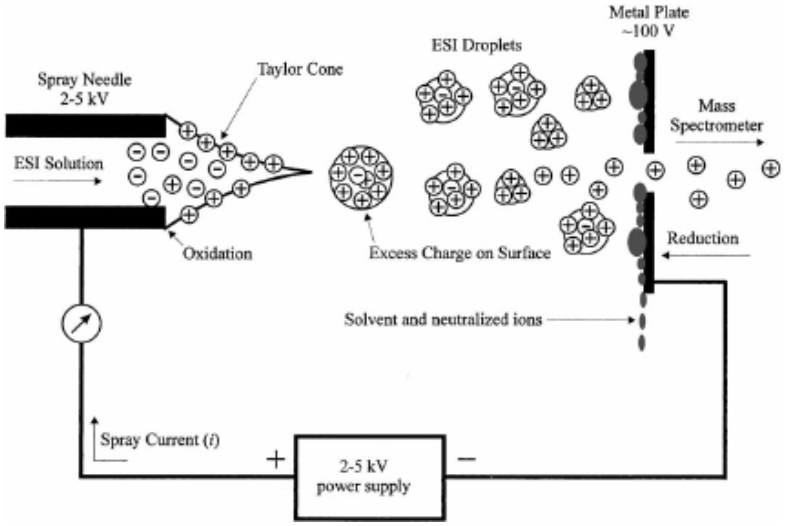


Figure 6: Schematic representation of ESI and charged droplets formation. The droplets reduce their size and are released as gas phase ions. Reprinted from K. Wanner *et al.*, [28].

1.3 Mass analyzers

1.3.1 Quadrupole

A Quandrupole (Q) mass analyzer [40] is formed of 4 electrodes with circular or hyperbolic cross section. The mass filtering operation is obtained by the application of a combined direct current (dc) and alternative current (ac) potential applied between the two opposite electrically connected rods [41] (see **Figure 7**).

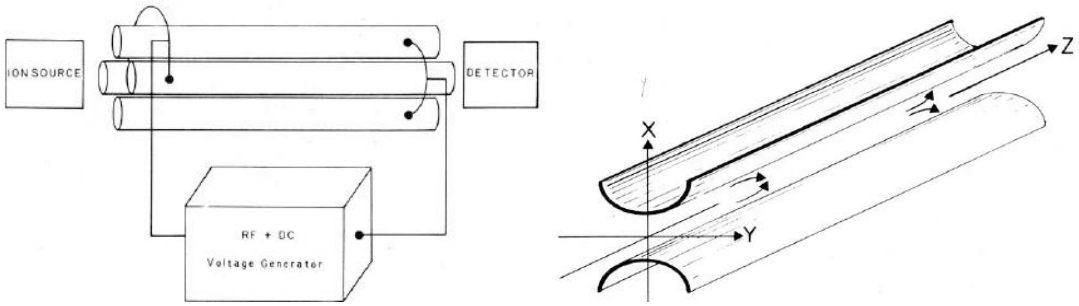


Figure 7: Schema of quadrupole and the trajectory of ions inside a quadrupole. Reprinted from P. E. Miller et al., [41].

The trajectory of ions in the quadrupole is explained by the influence of the applied potential ($\Phi_0 = +(U - V\cos\omega t)$ *and* $-\Phi_0 = -(U - V\cos\omega t)$) on charged molecules in the *x-z* ($F_x = m\frac{d^2x}{dt^2} = -ze\frac{\partial\Phi}{\partial x}$) and *y-z* planes ($F_y = m\frac{d^2y}{dt^2} = -ze\frac{\partial\Phi}{\partial y}$). In the *x-z* plane, positive ions are accelerated or focused in the centre of the *z-axis* during the positive phase of the ac waveform and defocused during the negative phase ($\Phi_{(x,\ y)} = \frac{\Phi_{0(x,\ y)}}{r_0^2} = \frac{(x^2 - y^2)(U - V\cos\omega t)}{r_0^2}$, with r_0 = radius of stable ions). As the dc potential is positive in the *x-z* plane, the average potential is mostly positive. The filtration operation of the quadrupole in the *x-z* plane is a high pass mass filter (in the case of positively charged ions) because light ions will tend to be influenced by the negative phase of ac waveform and discharged onto the electrodes ($a_{x,y} = \frac{8zeU}{m\omega^2 r_0^2}$ and $q_x = \frac{4zeV}{m\omega^2 r_0^2}$). In the *y-z* plane, the average potential is negative due to negative dc potential, and only light masses are focused in the centre of *z-axis* while heavy mass ions are discharged onto the electrodes. The

filtration operation in the *y-z* plane is said low mass filter (in the case of positively charged ions). A diagram of filtration of both axes is shown in **Figure 8.** Only ions belonging to the intercepted area can travel through the quadrupole without being deflected. A typical quadrupole mass analyzer operates at unit mass resolution, sufficient to separate two peaks one mass unit apart. The limit of *m/z* ratio detection is typically around 4000 [42].

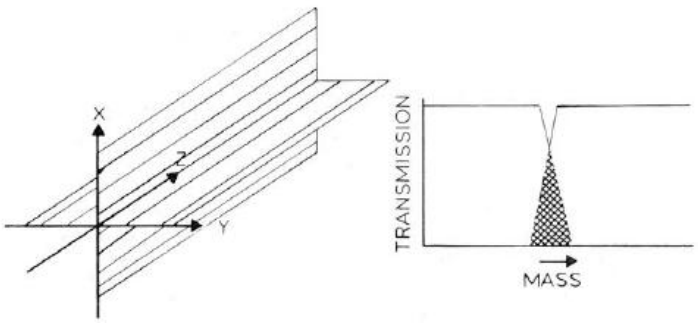


Figure 8: Superposition of *x-z* and *y-z* planes and area of selected *m/z* of ions. Reprinted from P. E. Miller *et al.,* [41].

1.3.2 Quadrupole ion trap

A quadrupole ion trap (QIT) [43][44] acts as a confinement device where ions are stored, formed by a ring electrode, and 2 end-cap electrodes (see **Figure 9**).

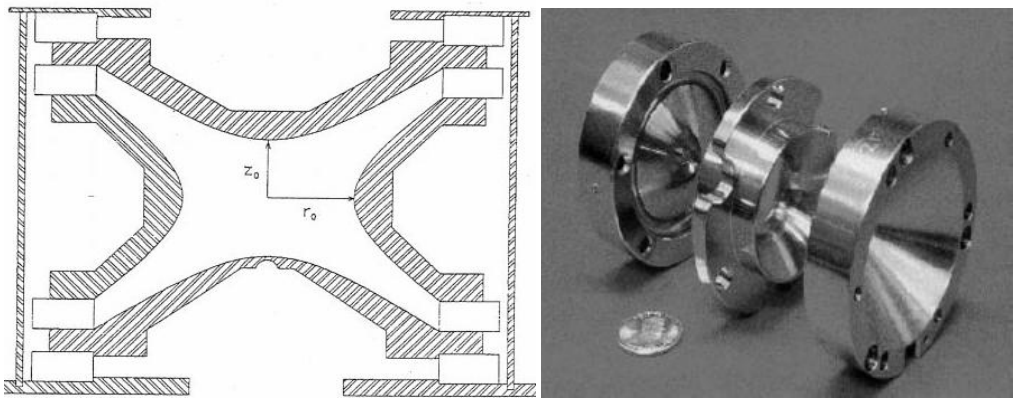


Figure 9: The composition of an ion trap. On the left, a schematic representation of an IT is shown with the axial (z) and radial (r) components. On the right, the picture represents the three electrodes that compose the trap. Reprinted from R. E. March *et al.*, [45].

As it was the case with quadrupole mass analyzer, a combination of ac and dc potential is applied to the circular and end-cap electrodes. However, in most commercial instrument the RF potential is only applied to the ring electrodes. In both cases ions are confined in the centre of the trap under a 3D quadrupolar field, described by its radial (r) and axial (z) components ($\beta_u = \left[a_u + \left(q_u^2/2\right)\right]^{1/2}$, $\frac{d^2z}{dt^2} + (a_{u-2q_u}\cos 2\xi)\mathrm{u} = 0$). Ions under the influence of this field are characterized by their axial and radial frequency, commonly called secular frequencies ($f_z = \beta_z v/2, \overline{D}_z = q_z\frac{v}{8} = \frac{ev^3}{m(m_0^r + 2z_0^2)\,\omega^2}$). The trap is filled with helium gas (dumping gas) that helps to stabilize ions via elastic collisions. The trajectory described by an ion in an ion trap looks like a flattened boomerang in the space between the electrodes (see **Figure 10**). A mass spectrum is generated by increasing the amplitude of the RF potential on the ring electrode. With increasing RF magnitude, low mass ions gain in oscillation amplitude and the ion cooling process is not sufficient to confine them in the trap. Thus, unstable ions are ejected by increasing order of *m/z* values. In QIT, the resolution depends on the scan speed. The relationship is inversely proportional. Therefore, a slower ejection time (expressed in $\frac{m/z}{s}$) of the trap will also result with a higher resolution. However, the resolution and mass accuracy in ion traps are also highly dependent of the total number of ions in it. With too many ions, saturation and space charging effects occurs. This effect creates a shield around the ions trapped in the centre. This modifies the action of the field and displaces the stability of ions. To overcome this effect, the number of ions in the trap has to be controlled. Space charge effect with 500 trapped ions was observed in QIT [29]. A typical QIT has a nominal mass range up to 5000 *m/z* units, a resolution of about 4000 FWHM, and an accuracy of about 100 ppm.

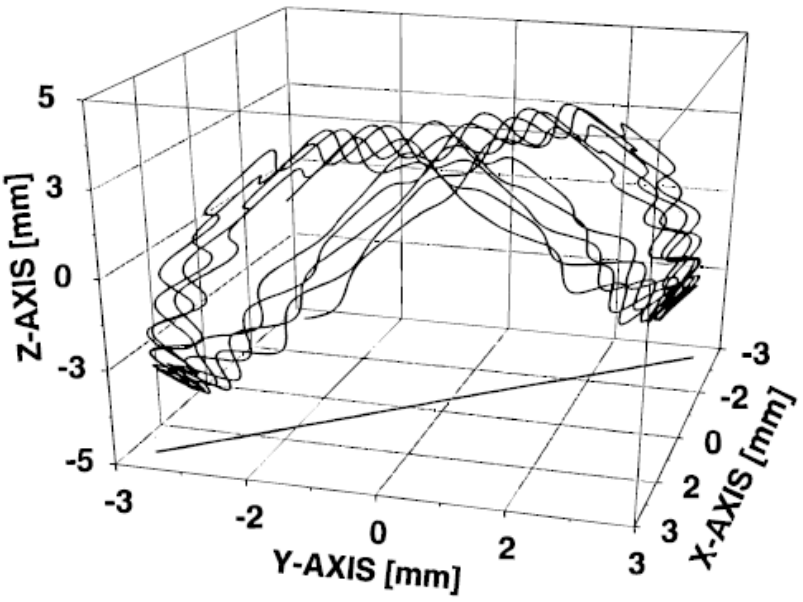


Figure 10: Ion trajectory of a certain m/z in the ion trap. Reprinted from R. E. March et al., [44].

1.3.3 Linear ion trap

Linear ion traps (LIT) are also known as 2D Ion Trap (2D IT). Physically, a LIT is a quadrupole device with end electrodes. In this trap, the ions are focused radially (*z-axis*) by the quadrupolar field. To store the ions in the trap axially (along the *z-axis*), a dc potential (positive for positive ions) is applied to both end electrodes. Several advantages can be drawn from LIT devices. The fact that ions are focused along the *z-axis* instead of a single point improves the sensitivity (more ions can be stored). Typically a LIT has more 10-fold higher ion trapping capacity than quadrupole ion trap. LIT with ion trapping capacity of 20'000 ions without space charge effect was reported [46]. Another advantage is that ions could be ejected radially or axially [47] by applying potentials to different electrodes. The mass range, resolution and accuracy of a LIT are similar to a QIT.

1.3.4 Time of Flight

The time of flight (ToF) mass analyzer separates ions according to their velocities. Ions in the flight tube are accelerated by a potential applied between an electrode and the extraction grid until they reach the field-free region. As all ions have the same initial kinetic energy, they can be separated during the flight and reach the detector according to their *m/z* (see **Figure 11**). An ion of mass *m* and charge *z* that leaves the ion source, accelerated by a potential V, and its electric potential energy E_{el} is converted into kinetic energy E_k: $E_k = \frac{mv^2}{2} = qV_s = zeV_s = E_{el}$. From this previous equation, the ion velocity is determined by: $v = (2zeV_s/m)^{1/2}$. Finally, the time needed to travel through the free-field region L (drift path) is given by $= \frac{L}{v}$. The ToF has no upper mass limit [48] and is a high-transmission efficiency mass analyzer (leading to high sensitivity) [49]. The analysis speed of a ToF is very fast (microsecond scale) but the mass spectrum contains a poor number of ions. Often, several mass spectra are recorded and summed to increase the number of detected ions. ToF mass analyzer can be used

in linear mode (LToF), with an electrostatic ion reflectron (RToF) [50] to increase mass-resolution and mass-accuracy and/or with orthogonal acceleration [51] of continuous ion beam to mimic a pulsed process. A modern ToF mass analyzer has a typical mass accuracy of 3-5 ppm and its mass resolution is 10'000 FWHM or higher.

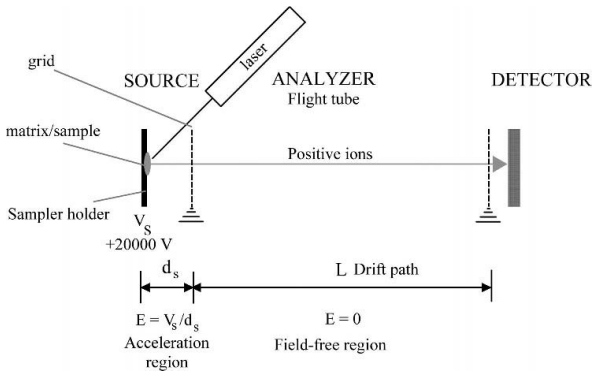


Figure 11: Time of flight ion detection schema. In ToF, the ions are separated in the free-field region. Ions from the source undergo acceleration due to a potential applied between the sources and the external grids (acceleration region). They keep the same trajectory through the drift path before hitting the detector. Reprinted from Hoffman et al., [29]

1.3.5 Ion Cyclotron Resonance

The ion cyclotron resonance (ICR) is a trap where ions are under the influence of an intense magnetic field (*B*), in order to have a curved or circular trajectory with a small radius (*r*). The stable trajectory of an ion in an cyclotron (cyclotron motion) (see Figure 12) is described by a balance between centripetal and centrifugal forces, respectively F and F' ($F = qvB$ and $F' = \frac{mv^2}{r}$). Thus, $qvB = \frac{mv^2}{r}$. The frequency of the ion is $\nu = \frac{v}{2\pi r}$ and consequently, the angular frequency is $\omega_c = 2\pi\nu = \frac{v}{r} = \frac{q}{m}B$. This equation shows that the ion frequency and angular velocity are independent of initial velocity of ions in the cyclotron. The ion mass is determined by measuring the frequency, and ions are expelled from the cyclotron by increasing the trajectory radius, thus their velocity. The ion detection is distinguished in two ways, 1) by resonance [52] and 2) by Fourier transform [53]. In the first case, an

electromagnetic wave of the same frequency of a given ion in the cyclotron is applied to induce an image-current of this ion, circulating near the electrostatic plates. This step is repeated for all targeted ions. In the second case, multiple ions are excited by a rapid phase over a large range of frequency. The complex wave detected as time-dependent function is transformed into frequency-dependent intensity function through a Fourier Transform to determine all ion's *m/z* in the cyclotron. ICR or FT-ICR can detect as few as 10 ions and up to 1000'000 ions. Its mass accuracy is about 1 to 5 ppm and its resolution up 100'000 (FWHM or higher).

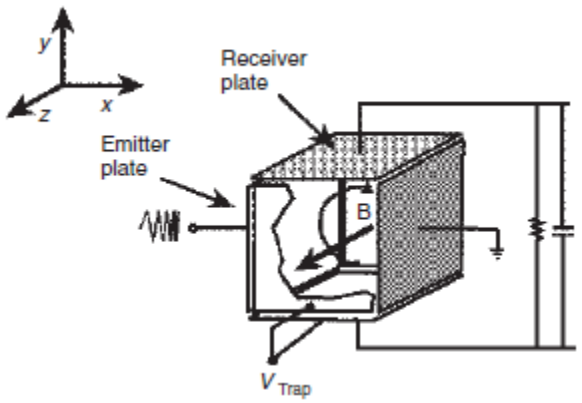


Figure 12: Diagram of Ion Cyclotron Resonance. Ions and magnetic field are oriented along the *z-axis*. As the magnetic field acts perpendicular to the ion velocity, the ion will undergo cyclotron motion along the *xy-axis* between electrostatic trapping plates. Reprinted from Hoffmann et al., [29]

1.3.6 Orbitrap

An Orbitrap is a mass analyzer composed of a central spindle-like inner electrode and an outer electrode with a barrel shape to generate an electrostatic field (Orbitrap) (see **Figure 13**). This mass analyzer was proposed by Makarov in 1999 [54], and further developments were made in 2000 [55] and 2006 [56]. The principle of an Orbitrap is simple. Due to the quadrologarithmic field, ions undergo axial harmonic motion and are trapped and orbiting around the inner electrode. The harmonic oscillating frequency of an ion along the *z-axis*

(around the inner electrode) depends on its *m/z* ratio and the potential between the electrodes. The angular frequency is $\omega = \sqrt{\frac{k}{m/q}}$. The frequency of the ions is measured by recording the induced current generated between two detector electrodes placed along the *z-axis* and a subsequent Fourier transformation of the measured signal. The Orbitrap has a high mass resolving power (100 000 FWHM or higher) and a mass accuracy lower than 3ppm.

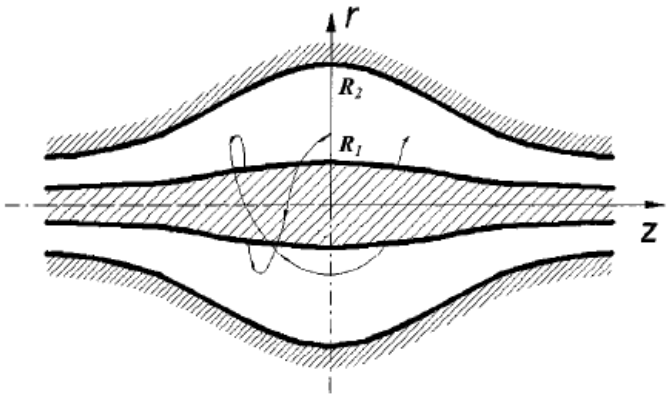


Figure 13: Schematic representation of a trap with stable ions trajectory (intricate spiral). Reprinted from A. Makarov et al., [55].

1.3.7 Hybrid mass analyzer

Hybrid mass spectrometry combines two or more different mass analyzers to have the advantage of both. A hybrid system is described by symbols of analyzers that ions travel through. One of the first hybrid system combined magnetic (B) or electric sector (E) to quadrupole (*e.g.* BEqQ or EBqQ) [57]. The advantage of this combination is to perform high-mass accuracy and resolution acquisition in the magnetic field sector and tandem mass spectrometry (described in the next chapter) acquisition in quadrupole sector. Due to high voltage applied in the magnetic sector, the kinetic energy of ions is increased and high collision energy activation of ions can also be performed in the collision cell (described in the next chapter). Other combinations are the coupling of ion trap analyzers with ToF [58] or ICR analyzers. It allows increasing the number of ions by their accumulation in the ion trap prior

to high-resolution mass analysis in the ToF or ICR analyzer. The combination of a ToF analyzer with a quadrupole is common for its simplicity, robustness, high-sensitivity, and performance [59]. In proteomics, the most used hybrid mass analyzer is the combination of an ion trap with an Orbitrap analyzer/detector. In particular, the linear ion trap is used in front of the Orbitrap for its sensitivity and selectivity of ions. The Orbitrap is simply used as a mass detector for its high-mass accuracy and resolution. Another particularity of an LIT-OT is also the ability of the instrument to perform parallel data acquisition in the LIT and Orbitrap. For example, during the acquisition of high mass resolving and accurate data in the Orbitrap, low resolution data could be obtained from the LIT by radial ejection for multi-stage mass spectrometry [55].

1.3.8 Tandem mass spectrometry

Tandem mass spectrometry (MS/MS) is a method involving at least two stages of mass analysis [60]. The procedure consists of a simple scan mode (MS^1) and product ion scans mode (MS^2). Typically, the mass spectrometer acquires a simple scan to get information on all molecular ion species. Then, the *n* most abundant precursor ions are selected for sequential isolation and activation in order to produce fragment ions and neutral fragments ($m_p^+ = m_f^+ + n$). The resulting product ions are then analyzed in a second stage of mass spectrometry. Tandem mass spectrometry analysis can be distinguished in two ways: tandem mass spectrometry performed “in space” by coupling two or more physical mass analyzers or “in time” by performing a sequential event of ion isolation and activation in the same mass analyzer. A typical tandem mass spectrometry experiment in space can be described by the operation of a triple quadrupole mass analyzer (QqQ). A QqQ is formed by three successive quadrupoles operated in different modes. Product ion scan illustrates well a typical use of a tandem mass spectrometer: an ion species with a certain *m/z* is filtered in the first quadrupole

(Q) and are sent into the second quadrupole (q) where the activation of ions occurs by collision induced dissociation (CID) with an inert gas. The second quadrupole is a collision cell. The collision cell act as ion guide and only ac potential (Radio Frequency, RF) is applied on the electrodes. The resulting ion products are sent into the last quadrupole to be filtered by their *m/z* ratio. The same scan mode can be performed in an ion trap to describe tandem mass spectrometry in time. In this case, each step of a product ion scan in a QqQ will correspond to an event taking place in the ion trap. So the first event will be the isolation of precursor ions, followed by CID experiment with the cooling gas already present in the trap (event 2), and product ions analysis (event 3). There is one particularity when tandem mass spectrometry is performed in an ion trap. At the moment when precursor ions are going to be activated with increased RF amplitude, a low-mass cut-off is created, preventing the trapping of ions which have *m/z* ratio lower than approximately one-third of the precursor *m/z* value. Tandem mass spectrometry is used for structural analysis of macromolecules such as proteins and peptides.

1.3.9 Collision induced dissociation

To perform tandem mass spectrometry, fragments from precursor ions have to be produced (ion activation). One of the activation methods is collisional activated dissociation (CAD) or collision induced dissociation (CID)**.** The principle of CID is simple; ions are accelerated and collide with a neutral gas. As the collision is inelastic, a few part of the available kinetic energy is transmitted into the ions as internal energy. The excess of energy in the ions result with their activation and subsequent dissociation. Harrison and co-workers have measured the conversion of kinetic energy into internal energy during the process of CID with n-butylbenzene precursor ion activation [61]. Depending on the magnitude of energy, the ion fragmentation pathway is different. Usually two collision regimes are distinguished. 1) High-energy collision, where Helium is used as standard gas and where the ion's excitation energy

is mostly due to the application of an electric field [62]. For high-energy collision, ions have to gain in velocity before colliding with inert gas. This energy can be provided by an intense magnetic or electric field (*e.g.* ICR) that accelerates ions before their activation in the collision cell (q). An electric field sector can be added before the collision cell to control ion velocity. 2) In low-energy collision, the collision gas is more important than the ion velocity due to conversion of kinetic energy into internal energy. A heavier and inert gas is preferred for an efficient translation of kinetic energy into internal energy. Argon, xenon or krypton are best choices of collision gas. The ion's excitation energy is mostly vibrational [63].

The inconveniences with CID methods are slow reaction rate of fragmentation for large molecules and the compromising of the vacuum system by the introduction of a collision gas. To overcome these limitations, other activation methods exist, such as surface-induced dissociation (SID) [64][65] where ions are accelerated against a metallic plate to induce fragmentation by collision. The pattern of fragmentation depends mainly on the nature of the ion, the collision energy and the type of surface. Methods without collisions, such as laser beam photodissociation or infrared multiphoton dissociation (IRPMD) were also developed [66]. Electron transfer dissociation (ETD) and electron capture dissociation (ECD) [67] are nowadays the most popular alternative activation methods. With ECD, multiply and positively charged molecules are bombarded with low-energy electrons emitted from a cathode. The irradiation of positive ions with low-energy electrons allows their capture with charge state reduction and the production of radical positive ions ($M^{+\cdot}$), which are in turn subjected to dissociation. There is dissociation because the recombination of energy after electron capture induces an increase of internal energy [67]. Generally, ECD is hardly used with 2D and 3D ion traps and quadrupole due to the strong electric field applied in it. However, electron transfer dissociation (ETD) is used in ion trapping instruments. It uses gas-phase ion/ion

chemistry to transfer an electron from singly charged anthracene anions to multiply charged molecules [68].

1.3.10 Protein and peptide identifications by mass spectrometry

The hyphenation of liquid chromatographs and tandem mass spectrometer, combined with automated data acquisition schemes made routine and massive protein identification possible. Two developments popularized the use of mass spectrometry in proteomics for protein identifications. First, the sample complexity is decreased by liquid chromatography to optimize molecular ion observation for tandem mass spectrometry. The sensitivity and selectivity of analytes can be controlled by combining different stationary and/or mobile phases, and by using an isocratic or gradient elution of analytes. The other advantage of using LC-MS/MS is its automation for a systematic analysis of biological samples. The second is the development of bioinformatics dedicated to mass spectrometry analysis. For example, automatic interpretation of tandem mass spectra, identification and quantification of molecular ions and related statistics to validate the results can be performed with suits of dedicated algorithms. With modern LC-MS/MS systems, a single LC-MS/MS analysis can capture the complexity of a complex mixture, such as cultured cells or a biofluid. The dynamic of such a mixture can only be observed in relative manner because the mass spectrometry analysis is affected by matrix effect and ionization suppression. For this purpose, isotopic dilution methods were developed to compare analytes and their relative quantifications (see chapter **protein and peptide quantification**). Finally, mass spectrometry offers many modes of operation to sample ions. Various ion sampling methods are used in routine to identify and quantify proteins or to target specific ion reactions. Many others are in work in progress with the development of new mass spectrometers.

1.3.11 Peptides fragmentation

In order to get structural information from mass spectrometry data, charged molecules must undergo dissociation. It can occur during ionization process but in practice, fragmentations during ESI or MALDI are usually neglected due to the low level of metastable ions. As previously described, ion activation consists in transferring enough energy to induce dissociation due to an excess of internal energy. Today, the most popular ion activation is CID due to its simple methodology and instrumentation.

Fragmentation of peptides can be placed in two categories: 1) cleavage along the peptide chain and 2) a cleavage of lateral amino acid chains [69][70]. In the first case, cleavage occurs at three different positions of a peptide bond chain: Cα-C, C-N or N-Cα. If the charge is kept on the n-terminal side of the cleaved peptide bond, the fragments will be respectively noted as a_n, b_n and c_n and when the charge is kept on the c-terminal side, the fragments will be respectively noted as x_n, y_n and z_n (see **Figure 14**). Internal fragments are also observed after two or more peptide bond cleavages. Among them, immonium ions result from the cleavage of two adjacent peptide bonds.

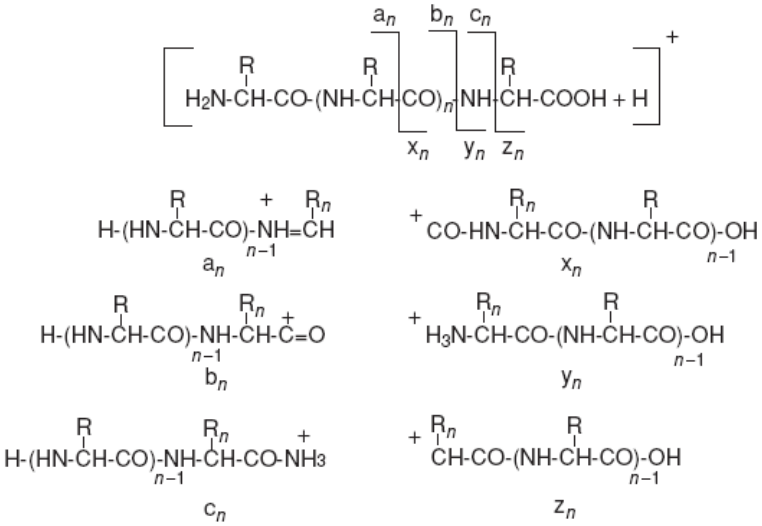


Figure 14: Theoretical peptides fragmentations in tandem CID mass spectrometry. Reprinted from E. de Hoffmann [29].

The second case, cleavages of lateral chains of amino acids, concerns only high-energy collisions. Such ion activation is useful to distinguish between Leucine and Isoleucine (isobaric amino acids). More generally, fragmentation patterns observed in a tandem mass spectrum are influenced by the position of the charges along the peptide backbone. The "mobile proton theory" describes these fragmentation patterns [71][72]. For example, protons are statistically more present on basic amino acids and charged amide groups, with more affinity to protons once in the gas phase, and thus considered as statistically favorable for dissociation. A summary of peptides fragmentation is shown in the **Figure 15** based on charge localization and the presence of basic amino acid in the peptide's sequence.

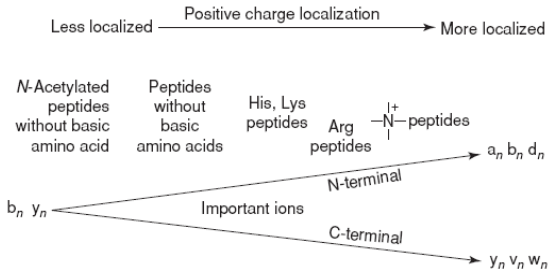


Figure 15: Charge and fragmentation with respect to the nature of peptides. Reprinted E. de Hoffmann [29]

1.3.12 Large-scale proteomics.

A proteome is defined as all proteins produced by an organism at a given time and condition [73]. The proteome is considered as more complex than the genome due to its dynamic complexity, *e.g.* post-translational modification and chemical modification of gene products, depending on the state organism or cells (*e.g.* cellular stress). Proteomics is the study of a proteome, a systematic analysis of proteins at large-scale to determine their structures, functions and condition-dependent variations [74]. In large-scale protein identifications, three strategies are used to analyze samples with mass spectrometry. Bottom-up, middle-down and top down proteomics. With the bottom-up approach, proteins are digested with trypsin prior to LC-MS/MS analysis. The resulting peptide fragmentation patterns are matched to peptide

sequences from protein sequence databases. The length of the peptide is dependent of the enzyme used for protein digestion. By extension, the principle of middle-down approach [75] was developed using digestion enzymes yielding longer peptide sequence compared to the bottom-up approach, such as the Glu-C enzyme. However, it is possible to directly analyze proteins without any sample preparation steps except ionization prior to tandem mass spectrometry analysis, the top-down approach. In contrast to bottom-up approach, the top-down approach is not affected by the non-reproducibility of enzymatic digestion and keeps the intact sequence of proteins with their post-translational modifications (PTMs). This opens the way of to study proteoforms [76]. Whatever the proteomic approach is, the resulting tandem mass spectra have to be identified by peptide spectrum matching (PSM). In summary, protein sequences are digested *in silico* and theoretical fragmentation pattern of peptides are generated. The experimental spectra are then matched by pair-wise comparison with theoretical spectra for identification [77]. The alternative way to interpret tandem mass spectra is to use the so called sample specific spectral libraries [78]. It uses confidently identified tandem mass spectra and builds consensus spectra by averaging tandem mass spectra that displays the same fragmentation pattern. Spectra in the spectral library are annotated, which means that all peaks in the spectra are known and contaminant peaks are discarded. Spectral library search is faster and gives better results than identifying peptides with a protein sequence database [79] [80]. Nevertheless, this strategy of identification is limited to already known peptide or protein sequences from previous mass spectrometry analyzes.

1.3.13 Protein quantification by mass spectrometry

Mass spectrometry is generally not considered as a direct quantitative technique due to ionization suppression and matrix effects. Thus, quantification is usually done via isotope-

dilution methods. In summary, an internal reference standard (IRS) is added for each compound to be quantified. The IRS is usually of same chemical structure, but with a different isotopic composition. The IRS has thus the same ionization response as the native compound, and relative quantification is performed by direct comparison of the ion abundances. For large-scale proteomics, a different isotopic label is introduced in most or all peptides/proteins from the sample, for example by growing cell lines on media with different isotopic composition, or by alkylating all peptide n-termini or cysteine-containing peptides with different isotopic label. In stable isotope labeling by amino acid in cell culture (SILAC) experiment, two populations of cells are grown in culture media that are identical. One population is fed with normal amino acids whereas the other one is fed with amino acids that incorporated C^{13}. Thus, by pooling extracted proteins from these two populations of cells, doublet of peaks corresponding to light and heavy labeled peptides are observed in simple scan mode. In opposition, and as the name suggests, label free quantification is a peptide/protein quantification method where peptides or proteins do not contain any label to be used as internal reference standard for isotope-dilution mass spectrometry. The samples are simply digested prior to LC-MS/MS analysis. Quantification is subsequently done by observing ion abundance variations between samples over several replicates [81]. The large number of replicates provides the accuracy and robustness of the quantification. The label-free approach has several advantages, such as minimum interventions for sample preparations, all present peptides can potentially be quantified, fast duty cycles, and no limitations in the number of samples to be compared [82]. Although it is generally assumed that isotope-dilution methods are more accurate [10], the technique is limited to the comparison of ten conditions per experiment so far [83] [84]. Label-free quantification is usually done in two ways: peak area measurements [85] and spectral count [86]. Spectral count is the easiest way of doing label-free quantification. It consists in counting the number of MS/MS spectra that

matches peptides of a protein. If a protein is more abundant in a particular sample, the mass spectrometer will select peptides more frequently, due to the limitations and redundancy of the data-dependent acquisition method. The number of peptide-spectrum matches (PSM) for this protein will thus be higher for this sample, indicating higher abundance. Several research articles describe spectral counting as robust quantification method and showed a good correlation with protein concentration and sensitivity to proteins abundance changes [87][88][89][90][91]. However, spectral count is dependent of the dynamic exclusion parameters; they should therefore be optimized prior to a spectral counting experiment. A balance between the number of *n* top abundant precursor ions selected for tandem mass spectrometry and the dynamic exclusion parameters should be found to affect as less as possible the sampling of ions and get as much as possible tandem mass data from a protein [92].

The peak area measurement methods extract each precursor ion abundance and chromatographic elution profile [93]. The calculated area under the curve (AUC) for each precursor ion is used to be correlated with peptide concentrations. The accuracy and sensitivity of extracted precursor ion chromatograms (XIC) are dependent of the *m/z* tolerance used (typically 5 to 500 ppm) for peak feature extraction. A peak feature is characterized by the ion's *m/z* ratio, AUC, and retention time. The features are extracted across the entire chromatographic elution, for each replicate. In addition, the chromatograms are aligned prior to condition-based comparisons. Integration of MS features and peptide identifications from PSM allows to measure peptide ratios between different samples. Intra- and inter-variances of proteins are measured and statistical tests such as t-tests are used to give confidence of differential protein concentration between different conditions [94][81][95][96]. Important steps in the label-free approach are the peak detection algorithm used to extract MS features, based on the measurement accuracy, and the resolution of the data for isotope pattern

detections and charge state determination [97][98]. Also important is the chromatogram alignment of the LC-MS data to be compared. Usually, the chromatogram alignment is done by taking into account ion *m/z*, retention time window and extracted precursor ion chromatograms. Often, a user-defined or automatically determined LC-MS analysis is used as reference for alignments [99]. Quality control (QC) standards can also be spiked in to improve the accuracy of the alignment. Andreev and co-worker proposed the notion of cross-assignments to integrate peptide identification data for alignments and showed improvement in terms of accuracy of quantification [100]. More generally, whatever the quantification methods is, either isotope-dilution chemistry or label-free based quantifications, the inherited issues from the DDA still remains a limiting factor for quantification.

1.4 Data acquisition strategies

Many different data acquisition strategies were developed to analyze molecular ions in mass spectrometers depending on the context of the study. In proteomics, three data acquisition strategies are mainly used during an LC-MS/MS experiment to identify and quantify peptides and proteins. The first one is commonly called data-dependent acquisition (DDA) and used for discovery proteomics. The second one is selected reaction monitoring (SRM) where specific reactions of targeted ions are monitored. The third and last one is data-independent acquisition (DIA), at the interface between discovery and targeted proteomics. The three methods will be described in the next chapters.

1.4.1 Data-dependent acquisition

In DDA, the *n* most abundant precursor ions are selected from a simple scan (MS^1) for subsequent product ion scans (MS^2) (see **Figure 16**). Thus, the tandem mass data acquisition is dependent of the previous simple scan for ion *m/z* and abundance [101]. The collected

tandem mass spectrometry data represents the peptide fragmentation patterns that will be used for peptide spectrum matching during the identification process. As the mass spectrometer selects only top abundant precursor ions, a bias towards abundant peptides is observed over chromatographic retention time. The problem of repetitive selection of the same precursor ions is resolved by using an exclusion list or by dynamically excluding ions from re-selection for a given time [24]. However, DDA still remains biased towards abundant peptides, preventing the selection of low abundant precursor ions, and thus limiting the dynamic range of identifications [102]. Others issues are semi-random selection of ions leading to a lack of reproducibility of DDA data [90] and the presence of chimeric or multiplexed spectra. These spectra are due to the co-isolation of precursor ion species for ion activation (also called "accidental CID" events). This leads to the acquisition of tandem mass spectra containing fragmentation pattern from more than one precursor ion.

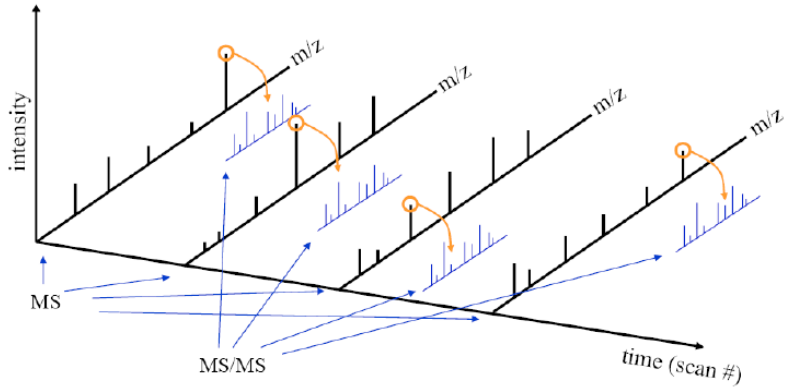


Figure 16: 3D representation of DDA mode. Each simple scan (MS^1) is followed by a dependent product ion scan (MS^2). Kindly from Markus Müller

To circumvent DDA issues, classical protein fractionation methods can be applied prior to LC-MS/MS analysis. For example, abundant proteins in a sample such as albumin in human plasma can be depleted by immuno-affinity chromatography [103]. Other protein and peptide separation methods such as ion exchange chromatography [104] or other methods are also

used based on the properties of the peptide mixture. For example, isoelectric focusing (IEF) is used to separate peptides or proteins according to their isoelectric point (pI) to decrease sample complexity prior to mass spectrometry analysis [105]. In any cases, DDA remains the most used mode due to its simplicity and user friendly operation. A non-experimented user can easily perform a LC-MS/MS analysis without any knowledge in mass spectrometry. Generally, DDA is sufficient to perform common discovery proteomics experiments [106]. However, a more selective and sensitive method is needed to detect and quantify proteins and peptides present at low amounts. Thus, alternative modes of data acquisition are currently investigated and introduced.

1.4.2 Selected reaction monitoring

An ion reaction monitoring experiment involves two stages of mass analysis and consists in monitoring ion reactions with fixed mass transitions [60]. The data representation of ion reaction monitoring experiments has no mass dimension and contains only the information of the mass transition abundances over retention time for a selected ion reaction. For each transition, a chromatographic trace is then calculated for the integration of the area under the curve (AUC) of measured ion reactions. Triple quadrupole (QqQ) mass spectrometers are the "gold standard" instruments for the SRM experiments due to their high sensitivity and selectivity to measure transitions in a very fast manner. A modern instrument can measure tens of transitions per second at highest sensitivity. This fast cycle time leads to the acquisition of enough data points over a chromatographic peak to fit a Gaussian elution profile of monitored ions. Another stage of ion activation can be added to confirm and validate the traces as in SRM cubed [107]. SRM is also implemented in ion trap (IT) mass spectrometers and is called pseudo-SRM (p-SRM) because transitions are extracted *in silico* post acquisition (see **Figure 17**) [108][109]. P-SRM offers the possibility to increase

precursor ion selectivity and to monitor an unlimited number of transitions per precursor ion [17] because the mass transition is not fixed for the second stage of MS. However, it should be kept in mind that SRM is a targeted analysis of already known/identified compounds, and it cannot be used for discovery experiments. The limitation of the SRM method is the number of transitions monitored per analysis. This number can be increased with the use of scheduled SRM [110], where target ion retention times are known and the mass spectrometer is configured to monitor only specific transitions during a given retention time window. SRM has become very popular during the last few years for proteomics and metabolomics applications [111][112][113].

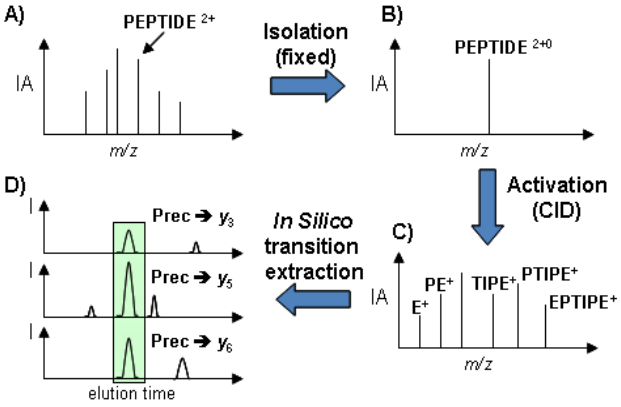


Figure 17: Principle of p-SRM acquisition in an ion trapping device. A) A complex mixture is infused into the mass spectrometer. B) Isolation of targeted peptide among different compounds. C) The targeted peptide is activated using CID and continuous product ion scans are recorded. D) p-SRM transition are re-assembled *in silico* after the complete dataset is acquired. Kindly from Alexander Scherl.

1.4.3 Data-independent acquisition

Data-independent acquisition (DIA) consist in acquiring tandem mass data independent from precursor ion abundance, in contrast to DDA, and integrates all precursor ions for tandem mass spectrometry analysis, in contrast to SRM (see **Figure 18**). For this purpose, the mass spectrometer is operated in parallel-CID mode [114][115][116][117][13][118][119]. With

parallel CID, the isolation window is larger than what one would typically use for DDA or SRM. If a complex mixture is analyzed, multiple precursor ions are selected and activated in parallel. The list of isolation windows (or ion channels [15]) is predefined and the windows are generally consecutive and incremental to cover the full precursor ion mass range. The mass spectrometer repetitively cycles through this list until the end of the analysis, and no simple scans are acquired in-between. The resulting tandem mass spectra consist of product ions from multiple precursor ions. In addition, the precursor-fragment ion lineage is lost, *i.e.* it is impossible to tell from a single tandem mass spectrum what fragment ion belongs to what precursor ion. The spectra are thus more complex compared to DDA acquisition. However, the method has several advantages. The frequency of tandem mass data acquisition is increased to improve the number of peptide/protein identifications (IDs) and their relative quantification compared to DDA [120]. In addition, the systematic interrogation of all precursor ion channels for tandem mass spectrometry is independent from precursor ion abundance. Consequently, low abundance ions can be identified [121]. Usually the width of ion channels and the frequency of the tandem mass data acquisition are the main factors that influence the quality and speed of data-independent mass spectrometry. A small ion channel of two to four *m/z* units is optimum to mimic DDA and to use existing peptide/protein identification platforms for data interpretation [15][122]. However, a large number of consecutive ion channels are necessary to cover the desired precursor ion *m/z* range, whereas ion chromatogram integrity has to be maintained as well. For this reason, multiple injections of the same sample are typically performed, to balance between the number of ion channels used per injection and the number of injections. In other words, if the desired precursor ion mass range is 400-1600 *m/z*, with an isolation window width of 2.5 *m/z*, and a cycle of 20 channels is used per LC-MS/MS analysis, 24 LC-MS/MS analyzes will be required to achieve the full analysis. If the channel width increases from 2.5 *m/z* to 10 *m/z*, only six injections of

the same samples will be necessary to cover the same precursor ion mass range. In contrast, with large ion channels (> 25 Th), one injection is enough to cover the *m/z* range of desired precursor ions. Such a method maximizes the duty cycle in order to preserve ion chromatogram integrity and minimize sample consumptions [123]. However, this comes at the cost of tandem mass data quality loss, increasing chimeric spectra rates and related difficulties to interpret correctly the acquired data without dedicated software for data processing. Other version of DIA maximizes the duty cycle by alternating between MS^1 and parallel-CID scans [13]. With these methods, all ions entering the mass spectrometer are activated. So theoretically, all ion information is recorded, and the best ion chromatogram integrity is maintained during the analysis. The complexity of data interpretation is partially overcome by high mass accuracy and resolution of acquired data for parent-fragment ion lineage [14][124]. Because the tandem mass data contains so many product ions per spectrum, it is difficult to identify what product ion peak corresponds to what precursor ion. Thanks to high mass accuracy and high resolution of scanned fragment ions, it is possible to discriminate product ions of a given precursor ion from others, comparing their single ion chromatogram shape and thus re-building parent-fragment ion lineage.

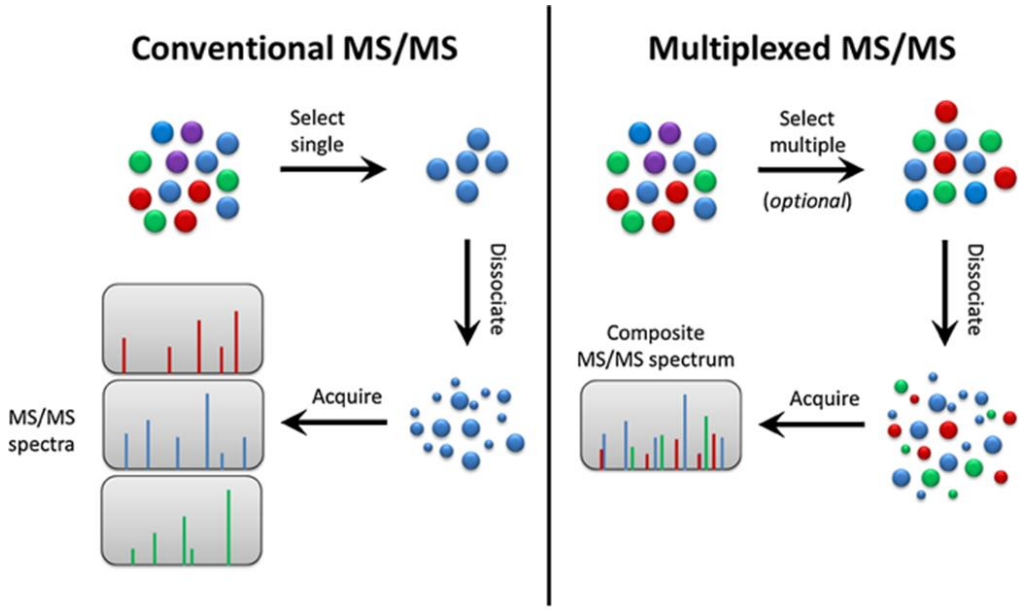


Figure 18: Comparison between data-dependent acquisition (conventional MS/MS) and data-independent acquisition (multiplexed MS/MS) methods. In DDA, precursor ions are sequentially isolated and activated. In DIA, all precursor ions within the isolation window are activated in parallel. The resulting tandem mass spectrum contains fragmentation patterns of all isolated precursor ions (composite MS/MS spectrum). Reprinted from J. D. Chapman *et al.*, [125]

1.5 Mass spectrometry data processing

Data processing is the interpretation of raw data generated by the instrument, generally as output result, into understandable objects for human beings or any other systems. It covers a large domain of sciences and especially computer sciences where (raw) data are translated, managed, and classified to be understandable and useable. In the field of mass spectrometry, data consists of information about the mass-to-charge ratio (*m/z*) and their abundance. Mass and charge state is derived from this data. With the addition of liquid chromatography for analyte separation, a third value corresponding to the observation of an ion at a given time (retention time) is associated to each molecular ion. Therefore, an ion is characterized by a three-dimensional coordinate: *m/z*, abundance, and retention time. The challenges for computer scientists involved in mass spectrometry-based protein data analysis is how to manipulate them without denaturing the structure in order to improve peptide/protein identifications and quantifications.

A typical DIA or DDA bottom-up proteomic experiment with LC-MS/MS consists in fragmenting ionized peptides within a user-defined isolation window. The resulting tandem mass spectrum does not contain only fragment ions of isolated precursor ion species but also other co-isolated molecules, commonly called contaminants or "noise peaks"[126][127]. The origin of these peaks is either from biological contaminants in the sample, chemical contaminants from solvents, and electric noise from instrumentation. One of the first challenges consists in removing these noise peaks. This has a direct effect on the calculation of the signal-to noise ratio (S/N). A low S/N indicates poor quality of the spectrum, where noise peaks are in competition with signal peaks specific to the observed precursor ions. Presence of noise can lead to a miss-assignment of a peptide spectrum match or a match being rejected during the statistical validation process. Several strategies exist to circumvent the noise peak issues:

- The first solution is to remove all peaks that have abundances lower than a user-defined cut-off value and assume that the loss of a few real peaks are negligible compared to the loss of myriads of noise peaks with low abundances. This simple way of processing spectra can drastically increase S/N and is commonly used in the mass spectrometry community. This way of processing is also called "top-hat" peak filtering [128].

- A more sophisticated way of removing noise peaks is to apply a peak detection algorithm in order to distinguish real peaks with isotope patterns (set of peaks related to ions with the same chemical formula but containing different isotopes) from electrical noise and chemical contaminants [129][98]. The resulting tandem mass spectra contain only peaks considered as real peaks by the peak detection algorithm.

Removal of peaks by top-hat processing or by peak detection approaches cleans spectra, reduces the number of product ions that have to be matched for molecular ion identification,

and increases S/N of mass spectrum. Many published algorithms for processing tandem mass spectra data combine both described approaches as pre-filtering or preprocessing [130] steps prior to the main data processing algorithm [85]. Listgarten and co-workers described an exhaustive list of algorithms for processing tandem mass spectra. They consolidated all the processing methods in three steps: 1) low-level data processing, 2) mid-level data processing, and 3) high-level data processing. The first level consists of building data structure to manipulate the data, to filter peaks and subtract background to minimize noise. The second step is used to facilitate the comparison of datasets and includes data normalization, peak detection and quantification, alignment for comparison, and error models to understand the variability of measured data. The last step optimizes data identification and statistical issues when multiple datasets are compared for biomarker discovery. It includes significance testing and multiple tests for data validation. The necessity of data processing is obvious: In a typical proteomics experiment, only 10% to 20% of the acquired data result with PSMs, which means that the remaining 80% to 90% of all tandem mass data is unused [120]. These high numbers are due to several reasons: low S/N, incomplete fragmentation pattern, contaminants, and the presence of post-translational modifications (protein sequences that can differ from the mere translation of the corresponding gene). Additionally, up to 30% of identified peptides can vary between technical replicates [90]. Mass spectrometry data processing increases protein identifications, reveals low abundance and modified proteins, and improves the specificity of database search. In terms of protein quantifications, it allows more accurate alignments and protein profiles measures, creating error models to handle data variability.

1.5.1 Data-dependent acquisition

It has been reported that at least 10% or more of the acquired tandem mass spectra [131] from DDA experiments have miss-assigned precursor ion *m/z* values. Often, the reported value is

shifted by one or more isotopes of the targeted precursor ion [131][132]. Another reason for miss-assigned precursor ion is overlapping isotopic clusters from different precursor ions. The instrument targets one ion, but another precursor is co-isolated. The reported precursor mass corresponds only to one precursor ion, and the other is thus incorrect. One solution to overcome this issue is to apply a large precursor ion tolerance (*e.g.* 1.1 Da) for PSM even if the precursor ion is measured with high-resolution and high measured mass accuracy [133][134]. Scherl and co-workers described a simple and efficient way to correctly assign miss-assigned precursor ion *m/z* from tandem mass spectrometry data by performing a peak detection in the region of precursor ion isolation, in the surrounding simple scans [131]. In cases of miss-assigned precursor ions, the reported values are corrected. In cases when more than one isotope pattern is present, the tandem mass spectrum is cloned and the value of each precursor ion is reported in one of the spectrum. This strategy leads to the identification of chimeric spectra and determination of their proportion in a typical DDA data set.

Chimeric or multiplexed tandem mass spectrometry data are characteristic of peptide co-elution during chromatographic separation [135]. Such events are frequently observed in DDA, and were described as “accidental CID” events. Precursor ions species with close *m/z* values are co-isolated and co-activated and this results with a tandem mass spectrum with product ions from more than one precursor ion. Chimeric tandem mass spectra are frequently miss-assigned during the PSM process due to the presence of intense fragment ion peaks from contaminant precursor ions [120]. In addition, the tolerance used for precursor ion candidates matching is often much lower as the width of the isolation window, resulting with unidentified spectra. One partial solution to this issue is to do multiple and iterative submissions of the same data to discriminate these spectra from other unique PSM, by counting and treating multiple and unique matches separately [136]. Houel and co-workers proposed software based on the detection of isotope signature of precursor ions from simple

scans in order to discriminate and classify chimeric spectra from DDA experiments. They reported an increase of false negative and false positive identification rates, leading to a decrease of identification sensitivity. In fact, no comprehensive solution is proposed by the scientific communities to solve the issue of chimeric spectra. Nevertheless, the problem is real and, unfortunately, common. It has indeed been estimated that more than 11 % of the acquired spectra from a complex sample are chimeric spectra [135]. The topic of chimeric spectra will be further discussed in the next chapter, in the context of DIA data processing, where the majority of acquired tandem mass spectra are the consequence of co-fragmentation events.

1.5.2 Data-independent acquisition

1.5.2.1 Multiplexed data

Multiplexed tandem mass spectrometry (MSX) consists in activating multiple precursor ions within a user-defined ion channel. It generates tandem mass spectra with multiple fragment-ions from multiple precursor ions. This strategy was initially used to overcome the low sampling rate of DDA for precursor ion selection. It was first described on a Fourier-Transform Ion Cyclotron resonance (FT-ICR) mass spectrometer. This technique took advantage of FT-ICR to retain a wide range of precursor ions for activation by stored waveform inverse Fourier Transform (SWIFT) excitation [137][138] or Hadamar Transforms (HT) [139][116] methods. After activation, product ion signals are re-linked to precursor ions through an encoding process based on excitation pulse events for precursor ion activation reactions and mass transfer pathways. The advantage was to acquire multiplexed tandem mass spectrometry data with high measured mass accuracy, resolution, and selectivity. The frequency of tandem mass data acquisition is virtually increased compared to serial data acquisition (DDA). The main difference between SWIFT and HT based multiplexed tandem

mass spectrometry is the number of spectra that have to be considered to produce one multiplexed spectrum. SWIFT is independent of the number of precursor ion dissociation but is dependent of resolution and *m/z* range. Masselon and colleagues described a data-dependent multiplexed approach where seven abundant precursor ions were retained using SWIFT waveforms and simultaneously dissociated by IRPMD [140]. The resulting tandem mass spectra were matched with a translated gene sequence database. They demonstrated the potential of using high-mass accuracy measurements for database search and subsequent peptide identification with multiplexed spectra. The pursuit of multiplexed tandem mass spectrometry strategy led to the development of the "Accurate Mass and Time" (AMT) Tags, consisting in storing for each identified molecular ion its accurate *m/z* and retention time [141]. This allowed to create a specific database with AMT features that kept growing with data accumulation from each experiment [142][143]. In contrast, HT is dependent of the number of activated precursor ions and independent of resolution and *m/z* range. Recently, HT based multiplexed approach was implemented on a hybrid quadrupole Orbitrap instrument. To mimic the combination of tandem mass spectra that must be acquired, five ion channels of four *m/z* units wide are randomly applied over a precursor ion mass range from 400-900 *m/z* [144]. A pre-defined and not overlapping list of 100 ion channels was used. Each multiplexed spectrum was considered as a linear combination of ion channels and fragment-ion abundance to build a system of linear equation solved by non-negative least square approximation [145]. It resulted with five de-multiplexed tandem mass spectrometry data per acquired spectrum. This multiplexing acquisition strategy combined with a post-acquisition de-multiplexing algorithm showed that interfering fragment ion signals were removed, which in turn increased S/N of tandem mass data. In terms of quantification, tandem mass spectrometry based quantification was less sensitive than simple scan based quantification. Nevertheless, the

demultiplexing processing approach was implemented in a popular open source software (Skyline) used for MS-based peptide quantification.

1.5.2.2 Shotgun-CID and MS^E data

Shotgun-CID [13] or **MS^E** [146] refers also to the idea of parallel activation rather than serial activation of ions as in DDA acquisition. In a proof-of-principle experiment, Purvine *et al.*, showed the feasibility of parallel-CID activation of ions in complex mixture of peptides and reported the possibility to reconstruct parent-fragment ion lineage from single ion current (SIC) chromatograms. Shogun-CID can be resumed to this simple but great idea of the reconstruction of parent-fragment ion lineage. The reconstruction is possible because precursor and fragment ions are measured with high mass accuracy , with almost 100% duty cycle. Simple scans are acquired with low in-source collision energy, and tandem mass scans are acquired at high collision energy (see **Figure 19**). Then, chromatographic elution profiles of all precursor and fragment ions are extracted. The reconstruction of parent-fragment ion lineage is based on the shape of elution profile of precursor and fragment ions. The experiment was performed with a simple ESI-TOF instrument, advantageous for high-mass accuracy and fast acquisition. However, the authors also described the implementation of the method in instruments with nominal resolution such as ion traps. A few years later, Silva and co-workers optimized the method by alternating the acquisition of precursor ion and fragment ions spectra (MS^E) during LC-MS/MS analysis in a Q-TOF instrument [14]. A full cycle of simple scan and MS^E scan acquisitions is achieved in four seconds, which preserved the chromatogram integrity. In parallel, the concept of Accurate Mass and Retention Times Pairs (AMRT) was introduced for a comprehensive analysis of such a data. The main attributes were mass, retention time, and metrics used to detected precursor ion chromatographic peaks for quantification (usually AUC). They also introduced an empiric quantification method based on the response of the three best ionizing peptides from a single protein [146]. A

quantification coefficient of variance less than 15 % was reported for protein quantification in complex samples. As mentioned above, shotgun-CID spectra cannot be used directly for identification due to the complexity to correlate fragment ions with precursor ions. Data processing using proprietary algorithms is thus necessary. Often, identification is based on previously acquired identifications to match the mass of precursor and fragment ions with high measured accuracy and retention times [147]. Because the chromatogram integrity is preserved, quantitative information can easily be extracted from identified peptides. Furthermore, pseudo-SRM or SRM assays can be conducted without limitations of limited number of transitions.

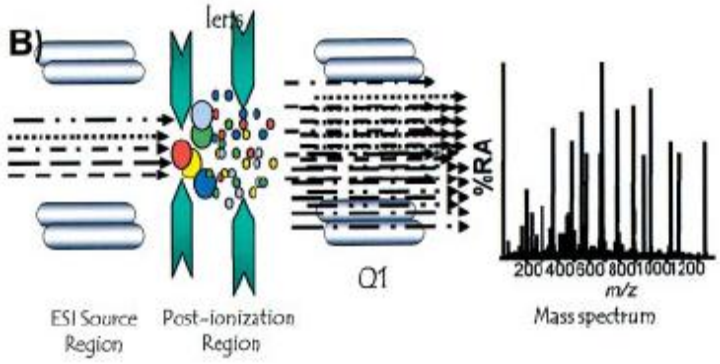


Figure 19: Schematic representation of shotgun-CID tandem mass spectrometry. Parallel-CID is performed by in-source fragmentation using the nozzle skimmer voltage in a quadrupole time of flight mass spectrometer. Reprinted from S. Purvine *et al.*, [13].

1.5.2.3 All ion fragmentation

All Ion Fragmentation (AIF) [124] is another name for a shotgun-CID acquisition method. The difference is that this method does not use an ESI-ToF but rather an Orbitrap hybrid system. The parallel CID activation of ions takes place in an octopole rather than in source (see **Figure 20**). Due to the high measured mass accuracy, resolution and fast acquisition frequency of such instruments, the overall performance is increased if compared to shotgun

CID in an ESI-ToF system. The data are processed in a similar way as for MS^E: MS^1 and AIF scans are deconvoluted, deisotopped and features are detected and integrated over retention time for AUC measurements [14]. The authors used the MaxQuant software to analyze the data (more details of this software will be provided in the next section). As for MS^E, good agreements between measured and expected spiked-in standards concentrations were reported for complex mixtures. The wishes of the authors was to show that AIF is as performant as MS^E, and that such a method can be employed in a variety of tandem mass spectrometers. This can potentially increase the use of shotgun-CID methods for proteomic experiments and its popularity.

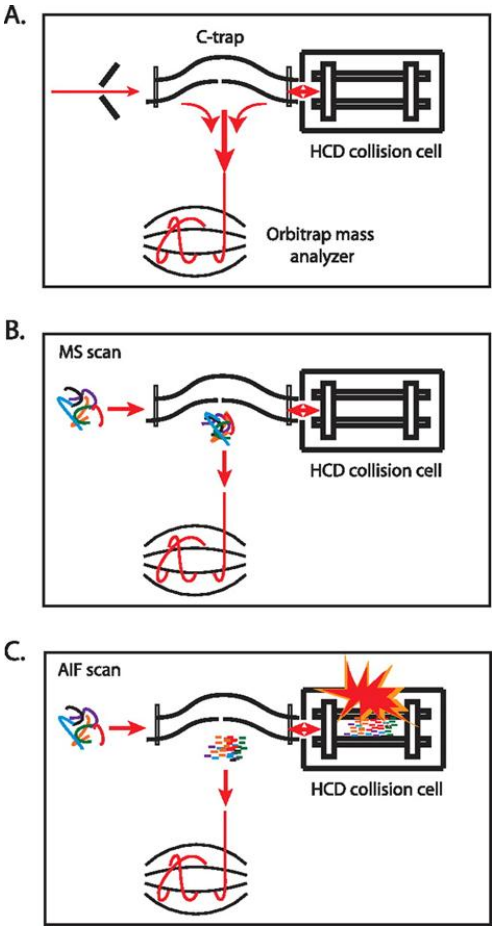


Figure 20: Principle of AIF. A) The mass spectrometer used for an AIF experiment contains a c-trap, a collision cell, and an orbitrap analyzer. B) For simple scans, the c-trap accumulates ions prior to Orbitrap analysis. C) For product ion scans, the packet of ions in the c-trap is activated in the HCD collision cell prior to Orbitrap analysis. Reprinted from T. Geiger *et al.*, [124].

So far, data-independent acquisition methods that maximize instrument duty cycle and preserve chromatographic integrity were described. Quantification studies on spiked-in standards reported good correlation between measured and expected protein concentrations. The complexity of data interpretation is overcome by using high mass accuracy measurements of both precursor and fragment ions. However, the challenge of parallel-CID activation is the correct assignment of fragment ions to precursor ions. For this purpose, DDA data were extensively used as templates for identifications and quantifications of shotgun-CID like data [147]. Other data acquisition strategies rely rather on complete sampling of precursor ions while maintaining a relatively narrow precursor ion isolation to limit data complexity. For example Venable and co-workers, used an isolation window of 10 *m/z* units where all precursor ions are activated for fragmentation [118]. The selectivity and sensitivity of tandem mass spectrometry data is increased due to the reduced number of precursor ions in such a relatively narrow isolation window, if compared to shotgun CID or MS^E. Consequently, the interpretation of tandem mass spectra is less complex. In addition, a simple scan was included before each cycle of tandem acquisitions. The simple scans were used to check the integrity of ion chromatograms and to detect precursor ions for data processing and subsequent peptide identification. Additionally, RelEx, an algorithm of correlation developed for proteins quantification [148] was applied to tandem mass spectrometry data for peptides quantification based on extracted fragment ion chromatograms. These chromatograms have improved S/N compared to base peak chromatograms. The inconvenient with such a method is that multiple injections of the same sample are mandatory to cover the full precursor ion *m/z* range. As the number of precursor ion windows is limited due to the acquisition frequencies of tandem mass spectrometers, multiple injections are indeed used. With each injection, other precursor *m/z* windows are used until the desired precursor *m/z* range is covered. Therefore, the necessary analysis time is increased.

1.5.2.4 Precursor ion independent from ion count

Precursor Acquisition Independent From Ion Count (PAcIFIC) can be considered as an extension of the DIA concept developed by Venable and co-workers, but with a more narrow isolation window, similar to the one used for data dependent acquisition [15]. This reduces the parallel-CID events to a minimum. The mass spectrometer acquisition cycle consists of ten tandem parallel-CID scans with an isolation window width of 2.5 *m/z* and an increment of the center of isolation window by 1.5 *m/z*. The overlap between consecutive isolation is 1 *m/z* minimizing the loss of isotopes for a given molecular ions. With PAcIFIC, the mass spectrometer acquires tandem mass scans from precursor ion mass range of 15 *m/z* for each analysis (see **Figure 21**). In the original publication, 67 injections were needed to cover the precursor ion mass range from 400 to 1400 *m/z*. The total analysis time was 4.2 days, with a relatively low frequency scanning mass spectrometer. In the meantime, improvements in ion trap scanning frequencies reduces the total analysis time by a factor 4, and this number is constantly improving with novel instrumentation. Despite the relatively long analysis time necessary for PAcIFIC, tandem mass spectrometry data can directly be submitted for identification in sequence database search engines developed for DDA. Also, due to the narrow ion isolation window and high sampling rate of PAcIFIC (each nominal *m/z* is tested for the presence of potential precursor ion), the total identifications are very high. Similar numbers are observed as with other fractionations methods, such as peptide isoelectric focusing, orthogonal chromatography or gas phase fractionation. Several studies have shown the ability of PAcIFIC to identify more peptides than other DDA-like methods [149]. During the data analysis, a comparable rate of chimeric spectra was found for DDA and DIA. The authors also reported identifications from tandem mass spectra where no corresponding precursor ion was observed in the simple scan, so-called orphan peptides. These orphan peptides are in agreement with the hypothesis that low abundance peptides are masked by

abundant peptides in the simple scans, and thus never selected for tandem mass spectrometry. More generally, PAcIFIC is a method dedicated for peptide identification rather than quantification. Indeed, the integrity of ion chromatograms for peptide quantitation is less preserved compared to shotgun-CID data acquisition with faster tandem mass sampling frequency. Nevertheless Panchaud and co-workers reported the ability of PAcIFIC to perform relative quantification using isotope-dilution methods with tandem mass tags [122]. The limitation of the low mass cut-off in the ion trap was overcome by pulsed-Q dissociation (PQD) to scan reporter ions in the trap at low *m/z* values [150][151]. As shown by Hengel and co-workers, spectral count seems also to be adapted for PAcIFIC because the peptide quantitation is directly related to peptide spectrum matches [149]. As reported by Panchaud and co-workers, PAcIFIC is limited by the scanning frequency of mass spectrometers. Higher acquisition frequency leads to better quantification. Indeed, an ion trap with a high scan rate capacity preserves the chromatogram's integrity, enlarges the number of precursor channels per cycle and shortens the sample analysis time. PAcIFIC data can be processed like any DIA data, with peak detection, denoising, and fragment ion chromatogram alignment for peptide quantification. However, as for any data-independent acquisition method, parent-fragment ion linage can be reconstituted in case of parallel fragmentation and precursor ion mass can be recalculated with more accuracy from complementary fragment ions [152].

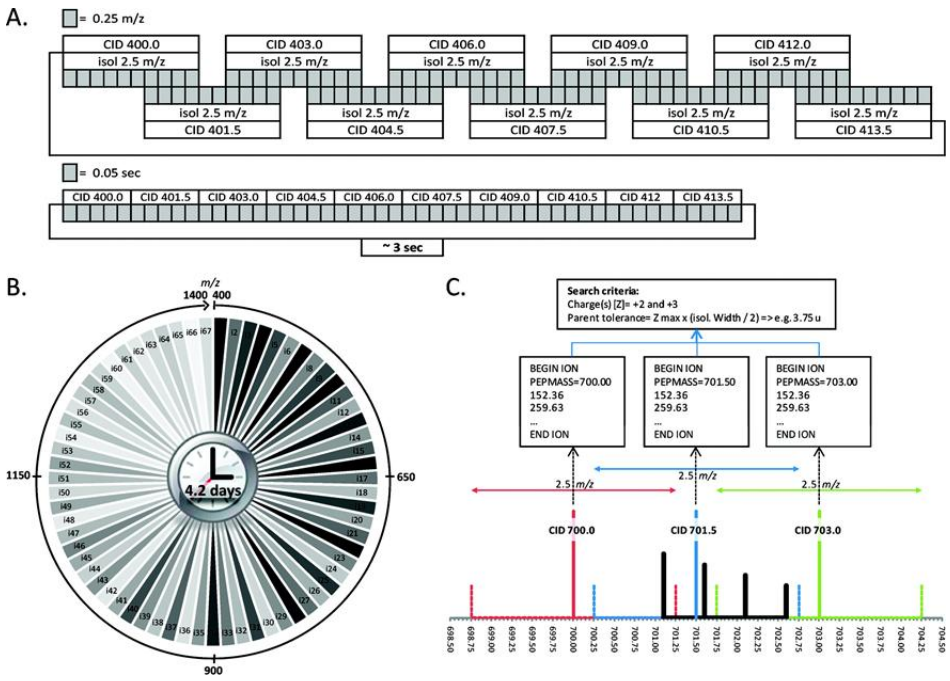


Figure 21: Principle of PAcIFIC. A) Pre-defined list of channels for one PAcIFIC LC-MS/MS analysis. B) To cover the desired precursor ion mass range of 1000 *m/z* (400-1400 *m/z*), 4.2 days of continuous analysis were necessary. C) Precursor ion mass tolerance used for database search identification. Reprinted from Panchaud *et al.*, [15].

1.5.2.5 XDIA

Extended data-independent acquisition (XDIA). This method combines electron transfer dissociation (ETD) and collision activated dissociation (CAD) of precursor ions for bottom-up and middle-down proteomics [153]. The authors have shown an improvement of 250% of peptide identifications with this combination. The novelty of this method is that peptide identification is not related on DDA data but on extracted ion signal from simple scans before each cycle of DIA tandem scan events. Specifically, a peak detection algorithm is used to extract potential precursor ions from simple scans [154]. Then, a set of precursor ions corresponding to the isolation window are assigned to the corresponding data-independent tandem mass spectra (see **Figure 22**). Additional filters are used to improve peptide identification, such as removing signals from neutral losses and charge reduced precursor ions. The resulting DDA-like tandem mass data is searched for peptide identifications. The advantage of XDIA is that the tandem mass data is identified with a narrow precursor ion

tolerance, improving specificity of database and/or library search. For peptide quantification, spectral count showed higher absolute numbers of PSM compared to DDA data, increasing the precision of the method. XDIA must be used with dedicated software, but an open-source XDIA processor is freely available. However, only tandem mass spectrometry data from an observed precursor ion can result with positive identification, whereas orphan peptides cannot be observed. It must be emphasized that PAcIFIC and XDIA both try to overcome the issues of undersampling and missing identification using systematic and narrow isolation windows or comprehensive peptide detection from simple scans.

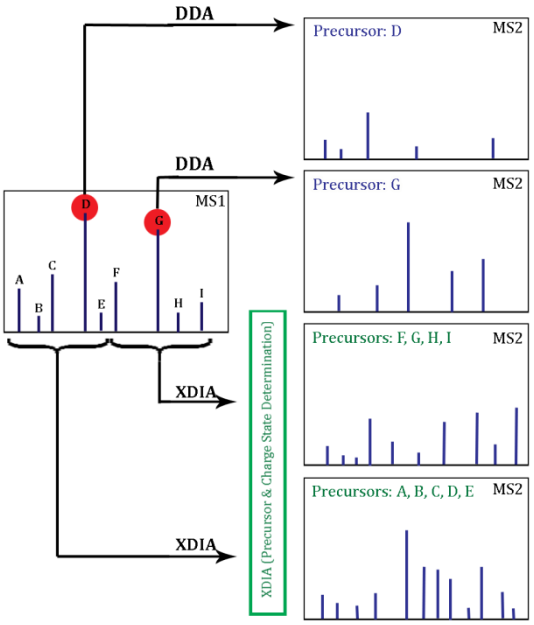


Figure 22: Data-dependent versus extended data-independent acquisition. In the DDA approach, a simple scan is acquired so precursor ions can be selected for posterior analysis by tandem mass spectrometry. In the XDIA approach, a wide precursor window including multiple peptide ions is selected for fragmentation. The precursors and their charge states are detected by the XDIA processor and are included in the final .MS2 file. Reprinted from P. C. Carvalho *et al.*, [153].

1.5.2.6 SWATH

Sequential window isolation of all theoretical mass spectra. SWATH is the acronym for this method [123]. Unlike shotgun-CID or PAcIFIC, SWATH uses an isolation window of 26

m/z with an overlap of 1 *m/z* between consecutive tandem acquisitions. The cycle of 32 SWATH scan events is used to cover a precursor ion mass range from 400-1200 *m/z*, which is topped by a simple scan. The full cycle, including the simple scan, is achieved in 3.2 seconds (see **Figure 23**). Chromatogram integrity is thus preserved if peak elution is 30 seconds or above. One single analysis is enough to cover precursor ions mass range of peptides from tryptic digests. As most DIA method, SWATH tandem mass spectrometry data cannot be directly submitted for identification. Sample specific spectral library have to be generated first from previously acquired data using another method. There are many advantages of using SWATH even though it cannot be used for discovery proteomics. First, precursor and fragment ions are detected with high measured mass accuracy and resolution in a Q-ToF MS instrument. Second, with the continuous growing of sample specific spectral libraries, most previously identified peptides using classical DDA acquisitions can be quantified without any additional data acquisition. Third, transition traces extraction is done as in multiple reactions monitoring high-resolution (MRM-HR) assay [155] with a resolution above 15'000 and accuracy below 20 ppm, ensuring highest specificity. Peptides can be quantified by systematic query of all fragment ion traces in the considered SWATH window [118][156]. These ion maps represent the signature of all fragment ions for each isolation window over chromatographic retention time. From this map, co-elution of product ion currents are observed, measured, and quantified. The authors reported identical quantification accuracy as compared to SRM in triple quadrupole instruments, but with limits of quantification about one order of magnitude higher. During these last two years, much work in proteomics and other *omics* fields used SWATH for peptide quantifications and biomarker researches [157][158]. The particularity of SWATH comes from the concept of acquiring simple scans and SWATH scans with high-resolution and high measured mass accuracy and a comprehensive data analysis workflow. The data from previously acquired sample spectral library is used to query

systematically transition for all observed peptides. In parallel, data processing and search times diminishes drastically [79]. If a spectral library of a full proteome is available, it is potentially possible to measure all proteins in a single SWATH-MS.

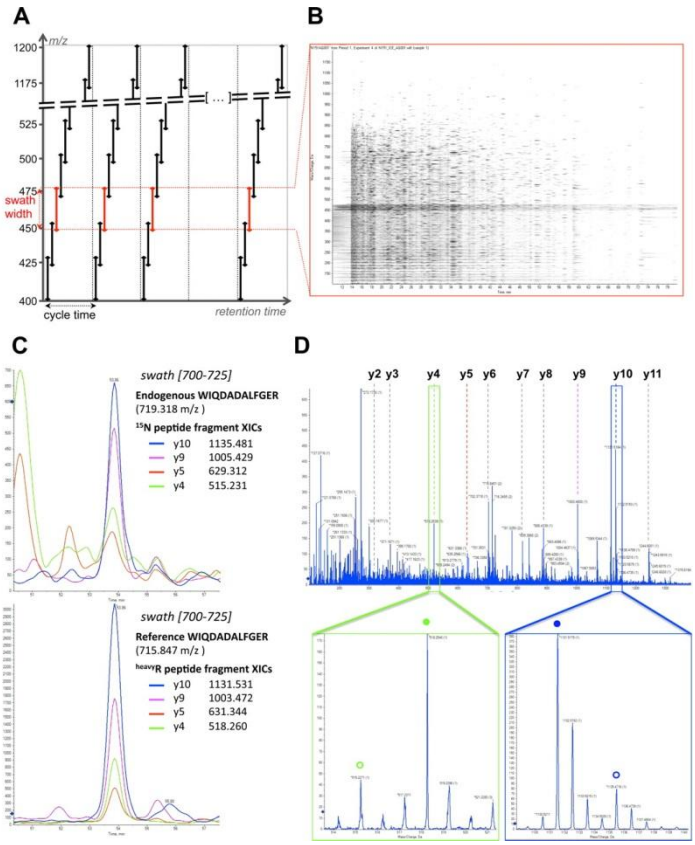


Figure 23: SWATH MS data-independent acquisition and targeted data analysis. A) Cycle of 32 discrete precursor isolation windows of 25 Th width. B) One swath tandem mass ion map. C) Targeted analysis from spectral library and chromatographic traces calculation. D) Display of correct extraction of fragment ions from step C to raw tandem mass data. Reprinted from L. C. Gillet *et al.*, [123].

1.6 Software and tools for multiplexed and data-independent acquisition

Among the large variety of software and tools available to process tandem mass spectrometry data, the ones particularly useful and/or specifically designed for data-independent acquisition are described below. A brief description is given for each of them. If possible, the type of tandem mass spectrometry data that can be processed with the software is indicated.

1.6.1 Extended data-independent acquisition Processor

Extended data-independent acquisition (XDIA) Processor. This software can be used for any type of data that includes a simple scan before each cycle of DIA tandem mass scans. The isolation window should not be too large in order to reduce the complexity of fragment ion patterns from parallel-CID activations. Thus, this software can process the original DIA, PAcIFIC, and SWATH data. As described by the authors, XDIA processor uses YADA (a subroutine) [154] that recognizes isotopic envelope pattern of peaks from two nearest simple scans. A list of precursor ions is then generated and combined with all tandem mass scans between the two simple scans. More precisely, if *n* precursor ions are detected within the list and fall into the isolation window of the tandem mass spectrum *i*, spectrum *i* will be written *n* times with *n* different precursor ion *m/z* values. Additionally, filters based on charge-reduced precursors and neutral loss ions are used to improve identifications of ETD data [159]. The resulting tandem mass spectrometry data is similar to DDA data with the exception of corrected precursor ion *m/z* ratios and inherited redundancy of the data. However, the redundancy of the data does not affects significantly the time of database search due to the use of a narrow precursor ion tolerance for peptide identifications.

1.6.2 Demux

Demux is a software to identify more than one PSM from DIA tandem mass spectra [160], such as after parallel CID events (chimeric spectra). The association of fragment ions and single molecular ion elution profile is done after deconvolution of tandem mass spectrometry data. Briefly, a matrix n x m, where n = number of *m/z* bins and m = number of interval of retention time, is generated from tandem mass scans that share the same isolation window. A clustering algorithm based on peak correlation is used with a sliding block of m = 100 with an

overlap of 50 %, to extract clusters of fragment ion peaks. For each cluster, a time-domain filter [100] is used to limit the considered molecular ion and fragment ion elution profiles to build synthetic tandem spectra. At this time, the consensus spectra are identified with a precursor ion tolerance of +/- 12 and 18 Da, respectively for 2+ and 3+ ions due to the large isolation window used (10 *m/z* units). Demux has an important capacity to reduce the data by clustering fragment ions to a user-defined sliding block, and reducing the time spent for database search. Additionally, peak binning and spectra averaging improves signal-to-noise ratio of synthetic tandem mass spectra. As observed even with most sophisticated processing algorithms, the use of large isolation window remains an issue. It might thus be interesting to use DeMux on narrow mass window DIA method such as PAcIFIC to test its capacity to simplify chimeric spectra.

1.6.3 Qcorr

QCorr is a tool developed to calculate ion molecular weight from tandem mass spectra by using cross-correlation functions [152]. The algorithm is based on the complementarities of fragment ions formed after cleavage of the same peptidic bond. Assuming that complementary ions are singly charged, their summation equals to the molecular weight of the peptide and two protons. In order to maximize cross-correlation of complementary ions, a tandem mass spectrum is correlated to its reversed spectrum within a user-defined isolation window. The accuracy of the re-calculated precursor ion mass is dependent of the user-defined peak binning value for tandem mass spectra comparison. The authors reported that QCorr performs well with +1 and +2 charged tandem mass spectrometry data but has limited performances with higher charge states due to the presence of contaminant peaks. DIA data were also tested with this algorithm, and improvement of the measured precursor ion accuracy was reported. The authors also noticed that isolation window width played a major role for the

correct calculation of the precursor ion mass. It should also be noticed that Qcorr was developed for low-resolution tandem mass spectrometry data. Thus, its use for narrow isolation window DIA methods such as PAcIFIC is convenient.

1.6.4 OpenSWATH

OpenSWATH is open source software dedicated to process SWATH-MS data [156]. The analysis of SWATH-MS data by OpenSWATH requires five major steps well described by Röst and co-workers. First, SWATH-MS data and assay library (sample spectral library) are converted into a suitable format for OpenSWATH. Second, openSWATH performs retention time alignment with correction from multiple chromatogram alignments. Third, thanks to the information contained in the assay library, fragment ion chromatograms are extracted with integrated ion abundance over retention time. Fourth, co-eluting ion chromatograms are grouped and scored with orthogonal scores during peak-group scoring. Last, statistical assessments are performed from step four for false discovery rate calculation of targeted assays. OpenSWATH is designed for a fully automated analysis of targeted and DIA mass spectrometry methods. Good accuracy of identifications with a pseudo receiver operator characteristic (ROC) > 0.9 and mean CVs less than 20% through all technical replicates was reported. The advantage of OpenSWATH is that it accepts several MS instrument manufacturer data format. It should be compatible with all type of DIA data for identification and quantification purposes.

1.6.5 MaxQuant

MaxQuant is an integrated suite of algorithms for MS-based proteomics data with high resolution and high measured mass accuracy, specially designed for stable isotope labeling

amino acids in cell culture (SILAC) based experiments [161]. The core of the platform is composed of: 1) Feature detection and quantification from simple scans, 2) improvement of peptide mass accuracy by using recalibration from standard spiked-in peptides, 3) peptide and protein identification using post error probability of each individual peptide, and 4) protein quantification using AUC integration over retention time. MaxQuant contains also additional packages like Andromeda and Perseus, for database search and statistical assessment of peptide identification and quantification, respectively [162]. In the context of DIA data processing, these suites of algorithms are very useful and MaxQuant was recently used to analyze AIF data. In short, 3D peaks detection was applied individually for simple scans and AIF data, and parent-fragment ion lineage was assessed on the basis of a cosine correlation [124]. The abundance of co-eluting fragment ions with precursor ions was summed to generate tandem mass data suitable for DDA database search. Despite this very sophisticated analysis of AIF data with MaxQuant, the challenging issue was that the resulting tandem mass data gave poor identification results. The method is thus probably not viable, at least in its present form, for discovery approaches. Probably, the same strategy of data analysis with a narrower isolation window for parallel-CID data should yield better peptide identification results. Even though, a loss in terms of quantification accuracy is expected with such an acquisition method.

1.6.6 Skyline

Skyline is a software dedicated for targeted and quantitative proteomics data analysis [163]. It can create and edit targeted proteomic experiments, support spectral library matching, and fully supports SRM quantification experiments. The window-based graphical user-interface and rich data interactions made Skyline very popular for method development and complex data displaying. Skyline uses the CRAWDAD peak detector to calculate transition peaks and

incorporates SSRCalc to predict theoretical peptide retention times based on hydrophobicity scores [164]. Most DIA data can benefit from the rich visualization tools offered by Skyline. This is especially true for large isolation window DIA, where the use of spectral library is almost mandatory for identification and quantification. SWATH-MS data are typically well suited for identification and quantification of peptides with Skyline. As described previously, MSX methods to de-multiplex tandem mass data acquired from multiplex data-independent acquisition methods is also implemented in Skyline [144]. Skyline generates an inclusion list of isolation windows for subsequent data acquisition. Once the data is acquired and submitted to Skyline, each multiplexed tandem mass spectrum is de-multiplexed and the results can be visualized and analyzed. Therefore, the effect of de-multiplexing at the spectrum and transition level is visualized. Spiked-in standard proteins were measured with Skyline, and good accuracy was reported. Skyline will probably be the most used graphical interface software for complex data processing and visualization in the future.

2 Label-free protein quantification on tandem mass spectra in an ion trapping device

This chapter describes label-free quantification based on tandem mass spectrometry ion abundance using LC-ESI-MS/MS analysis. We showed that selecting precursor ions for tandem mass spectrometry independent from their abundance can increase the dynamic range of identified molecular ions, and that their relative quantification is possible from such data.

The linear ion trap was operated in pseudo-multiple selected reaction monitoring (p-mSRM) mode, scanning repetitively a list of predefined precursor ion *m/z* during the chromatographic separation. For each precursor ion, tandem mass data was acquired regardless of its ion abundance and during its elution time. Ion chromatogram traces were extracted for each precursor ion and compared to label-free quantification using data-dependent acquisitions. A dynamic range of quantification of five orders of magnitude was reported with p-mSRM, and showing linearity with a spearman coefficient correlation of 0.99. With data-dependent acquisition, the dynamic range of quantification was limited to three orders of magnitude. We concluded that relative label-free quantification was possible for protein concentrations from 10 amol to 1 pmol, using data-independent tandem mass spectrometry, in a linear ion trap.

JIOMICS | VOL 1 | ISSUE 2 | DECEMBER 2011 | 211- 215

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i2.45

Label-free protein quantification on tandem mass spectra in an ion trapping device

HuiSong Pak[1], Carla Pasquarello[1], Alexander Scherl[1]*

[1]Proteomics Core Facility, Faculty of Medicine, University of Geneva, Geneva, Switzerland

Received: 27 October 2010 **Accepted:** 23 February 2011 **Available Online**: 14 March 2011

ABSTRACT

Label free quantification using liquid chromatography and electrospray ionization tandem mass spectrometry (LC-ESI-MS/MS) is widely used in quantitative proteomics. However, data-dependent bottom-up proteomics suffers from low reproducibility due to semi-random selection of precursor ions for tandem mass spectrometry. In addition, this acquisition mode is biased towards abundant peptides. To overcome these problems, alternative precursor-ion selection methods were developed, such as data-independent acquisition and *pseudo*-multiple selected reaction monitoring (p-mSRM). With these methods, several tandem mass spectra are acquired over the chromatographic elution time of precursor ions. In this report, we investigated if the acquired tandem mass spectra can be used for label-free quantification. For this, extracted fragment ion currents were correlated to relative protein concentration. A linear relationship between ion current and proteins concentration was observed over five orders of magnitude. Thus, we conclude that relative label-free peptide and protein quantification can be performed in an ion trap using the data-independent acquisition mode.

Keywords: *pseudo*-multiple selected reaction monitoring; quantitative analysis, ion trap, peptides.

Abbreviations

B-LG: β-Lactoglobulin; **CID:** Collision induced dissociation; **DDA:** Data-dependent acquisition; **DIA:** Data-independent acquisition; **LC-ESI-MS/MS:** Liquid chromatography electrospray tandem mass spectrometry; **mSRM:** Multiple selected reaction monitoring; **PAcIFIC:** Precursor Acquisition Independent From Ion Count; **p-mSRM:** *Pseudo*-multiple selected reaction monitoring; **SD:** Standard deviation; **SRM:** Selected reaction monitoring; **Trp II:** Trypsin inhibitor type II .

1. Introduction

Over the past years, direct interfacing of liquid chromatographs with tandem mass spectrometers (LC-MS/MS) has become a very popular tool for qualitative and quantitative analysis of complex peptide mixtures, such as enzymatic digestion product of complex protein samples. In particular, data-dependent acquisition (DDA), or on-the-fly precursor ion selection for isolation and subsequent activation and tandem mass analysis is widely used [1]. MS survey scans are acquired over the full mass range of peptide precursor ions and over the entire chromatographic elution time. The peptide precursor ion signals can then be used for label-free relative quantification [2,3]. With this quantification method, ion currents from identical peptides from different samples are directly compared to each other.

Although dynamic exclusion [4] prevents redundant acquisition of the most abundant peptides, DDA is still biased towards abundant species. In addition, the selection contains a random component limiting reproducibility of identified peptides [5,6]. Thus, a major effort is focused on developing alternative precursor-ion selection methods. With the so-called "Precursor Acquisition Independent from Ion Count" (PAcIFIC) or "Data-Independent

***Corresponding author:** Alexander Scherl. Proteomics Core Facility Centre Medical Universitaire Rue Michel-Servet 1 1211 Genève 4; tel: +41 22 379 5494 fax: +41 22 379 5926; E-mail Address: Alexander.Scherl@unige.ch

Acquisition" (DIA) mehods, continuous, non overlapping mass-to-charge ratio *(m/z)* windows are selected for isolation and subsequent activation [6,7]. Alternatively, no isolation is performed and all precursor ions are co-fragmented. The precursor-fragment ion lineage is then reconstituted post-acquisition using the chromatographic elution profile of all species [8,9]. A major advantage of these data-independent methods is that several tandem mass spectra are acquired over the entire chromatographic elution profile of all ions. Thus, fragment ion currents can be extracted for each precursor ion and used for quantification. Similarly, precursor-fragment ion transitions can be reconstructed post-acquisition, a term referred to as "*pseudo* -multiple selected reaction monitoring" or p-mSRM [10,11]. This mode of operation is described in Figure 1.

The term *"pseudo"* refers to the computer calculated post-acquisition reconstruction of the transition. Indeed, the full fragment ion spectrum is recorded here, by opposition to classical mSRM acquired in triple quadrupole instruments. *Pseudo*-mSRM has been showed to provide excellent quantitative measurement over a large dynamic range for small molecules [12] and peptides [13,14] in simple matrices and for non-complex mixtures. Here, we show that label-free MS2-based quantification is possible over at least five orders of magnitudes in very complex matrices, *e.g.* digested human plasma.

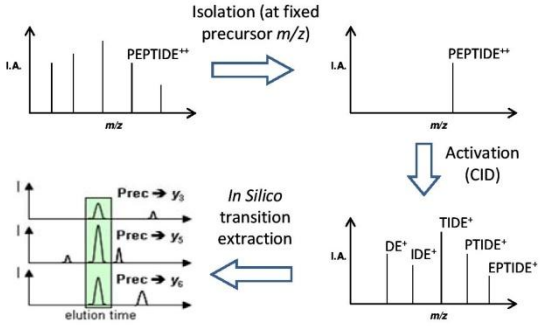


Figure 1. Principle of Selected Reaction Monitoring (SRM) and *pseudo-multiple* Selected Reaction Monitoring (p-mSRM). From a complex mixture of peptides, precursor ions are isolated and activated, typically by CID. All product ions are collected in the ion trap and scaned out according to their *m/z* ratio. This operation is repeated during the entire chromatographic elution of the peptides. Once all the data is collected, the ion currents corresponding to all desired transitions are reconstructed *in silico*.

2. Materials and methods

Liquid plasma, β-Lactoglobulin (B-LG), trypsin inhibitor type II (Trp II), iodoacetamide (IAA) and acetonitril (AcN) 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) was respectively from Promega (Madison, WI, USA) and Biosolve (Valkenswaard, the Netherlands). Stationary phases for columns were from Michrom (Auburn, CA). Analytical column (OD = 375 µm, ID = 75 µm, L=150 mm) and pre-column (OD = 375 µm, ID = 100 µm, L=20 mm) was made from fused silica tubing from BGB Analytik AG (Boeckten, Switzerland)

Human plasma and standard proteins were digested as previously reported [15]. In short, 500µg of B-LG and Trp II were dissolved in 200µl of 6M Urea and 50mM AB. 10µl of DTE 38mM was added and the solution was incubated at 37° C for 60 min (reduction). Then 20µl of IAA 108mM was added for alkylation during 60 min in the dark. Liquid digestion was performed overnight, by adding 25µl of trypsin (0.2µg/µl). The digested solution was desalted with a C18 micro-spin column (Harvard Apparatus, Holliston, MA, USA) and dried. In order to have aliquots of 10pmol/µl, dried solutions were dissolved in AcN/FA/H_2O 5/0.1/94.9%. These digests were spiked at various concentrations in trypsin digested, non-depleted human plasma. The concentrations were adjusted so that a total amount of 1, 10, 100 atomoles, 1, 10, 100 femtomoles and 1 picomole standard proteins were injected in our LC-MS/MS system. The injected plasma amount was kept constant of 0.126 µg per injection (corresponding approximately to one picomole of albumin).

The LC-MS/MS system consisted of a NanoAcquity chromatograph (Waters, Milford, MA) interfaced with an LTQ-Orbitrap velos mass spectrometer (Thermo Scientific, San Jose, CA). Peptides were trapped on a home-made, 20 mm long precolumn of 100 µm inner diameter and separated on a 150 mm analytical column of 75 µm inner diameter. The analytical separation was run for 65 min using a gradient of H_2O/FA 99.9%/0.1% (solvent A) and CH_3CN/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.

All samples were analyzed from the most diluted to the most concentrated, and new chromatographic columns were used for each technical replicate. The mass spectrometer was operated in the following conditions: for DDA, full MS spectra were acquired in the Orbitrap detector from *m/z* = 400 – 2000. The target ion population was 500,000 ions. Tandem mass spectra were acquired in a data-dependent manner in the linear ion trap on the five most abundant precursors (if present). Dynamic exclusion was set to one minute. Precursor isolation window was set to 2.0 *m/z* units. Normalized collision energy was set to 35%. For *p*-mSRM acquisitions, full MS spectra were acquired in the Orbitrap detector from *m/z* = 400 – 2000. The target ion population was 500,000 ions. Tandem mass spectra were acquired on the three β-lactoglobulin peptide precursor ions and on the three trypsin inhibitor peptide precursor ions listed in table 1, in a data independent manner. In other words, the precursor-ion *m/z* was isolated and fragmented over the full chromatographic analysis, no matter if precursor ions were present or not. The scan sequence was Full MS1 (orbitrap acquisition), CID of

JIOMICS | VOL 1 | ISSUE 2 | DECEMBER 2011 | 211- 215

m/z=858.406, CID of *m/z*=545.929, CID of *m/z*=623.294, CID of *m/z*=588.316, CID of *m/z*=600.858, CID of *m/z*=928.475. All CID spectra were acquired in the linear ion trap. The target ion population was set to 10,000 ions. The precursor isolation window was set to 2.0 *m/z* units. Normalized collision energy was set to 35%. All analysis were run in technical replicates (n=2). The first three DDA analyses, spiked with respectively 10 and 100 amol of standard proteins were used for database search.

Protein identification peak lists were generated from raw data using the embedded software from the instrument vendor (extract_MSN.exe). The monoisotopic masses of the selected precursor ions were corrected using an in-house written Perl script [16]. The corrected mgf files were searched against the SwissProt/Uniprot database (release 15.10 of 03-Nov-2009) using Phenyx (GeneBio, Geneva, Switzerland). Homo sapiens taxonomy was specified for database searching (34785 sequences) and the two standard protein sequences were added. The parent ion tolerance was set to 10 ppm. Variable amino acid modification was oxidized methionine. Fixed amino acid modification was carbamidomethylation of cysteins. Trypsin was selected as the enzyme, with one potential missed cleavage, and the normal cleavage mode was used. Only one search round was used with selection of "turbo" scoring. The peptide p value was 1 E-2 for LTQ-OT data. False-positive ratios were estimated using a reverse decoy database [17]. All datasets where searched once in the forward and once in the reverse database. Separate searches were used to keep the database size constant. Protein and peptide score were then set up to maintain the false positive peptide ratio below 1%. This resulted in a slight overestimation of the false-positive ratio [17]. For all analyses, only proteins matching two different peptide sequences were kept.

Quantitative data (extracted ion chromatograms and chromatographic peak integration) were extracted with Xcalibur 2.6 (Thermo Scientific). Base-to-base peak integration was performed manually using the "add peak" function of Xcalibur. The peak area value calculated by the software was used.

3. Results and discussion

In a first experiment, digested human plasma and the two digested standard proteins â-lactoglobulin (B-LG) and trypsin inhibitor type II (Trp II) were separately analyzed by DDA ESI-LC-MS/MS. This resulted with the identification of an average of 926 (SD=12) unique peptides from human plasma, corresponding to 413 (SD=18) unique proteins identified with at least two peptides. These numbers demonstrate the quality of the plasma digestion and the reproducibility of the data-dependent analysis. Then, for each standard protein the three peptides giving raise to the three most abundant precursor ion signals were selected for further analysis (Table 1). These 3 peptides correspond to sequence coverage of 15% for B-LG and 19% of Trp II.

In the next experiment, the digested proteins were mixed with digested plasma so that the injected amount varied from one atomole to one picomole. The quantity of plasma was kept constant, at 0.126 ug per injection (corresponding to aproximatively 1 pmol of injected albumin). The peptide mixture was then analyzed using DDA and *pseudo*-mSRM. For each concentration, the acquired MS1 spectra were inspected for the presence of the three peptide precursor of each standard protein. The isotopic cluster corresponding to peptide LFSNPTQLEEQCHI was visible at injected amounts of 1 femtomole and above. The two other peptides were dis-

Table 1. Observed peptides for MS1 and MS2-based quantification and their detection limit

Protein	Peptide	MS1 detection limit	p-SRM transition	MS2 detection limit	retention time (min)
β-Lactoglobulin	LFSNPTQLEEQCHI	1 fmol	$[M+2H]^{++} \rightarrow y_6^{+}$	100 amol	43,04
			$[M+2H]^{++} \rightarrow y_7^{+}$	100 amol	43,04
			$[M+2H]^{++} \rightarrow y_{10}^{+}$	10 amol	43,04
β-Lactoglobulin	TPEVDDEALEKFDK	10 fmol	$[M+3H]^{+++} \rightarrow y_{10}^{++}$	100 amol	33,97
			$[M+3H]^{+++} \rightarrow y_{11}^{++}$	100 amol	33,97
			$[M+3H]^{+++} \rightarrow y_{12}^{++}$	10 amol	33,97
β-Lactoglobulin	TPEVDDEALEK	10 fmol	$[M+2H]^{++} \rightarrow y_7^{+}$	100 amol	26,56
			$[M+2H]^{++} \rightarrow y_8^{+}$	100 amol	26,56
			$[M+2H]^{++} \rightarrow y_{10}^{++}$	100 amol	26,56
Trypsin Inhibitor	NELDKGIGTIISSPYR	1 fmol	$[M+3H]^{+++} \rightarrow y_5^{+}$	100 amol	39,98
			$[M+3H]^{+++} \rightarrow y_6^{+}$	100 amol	39,98
			$[M+3H]^{+++} \rightarrow y_{14}^{++}$	1 fmol	39,98
Trypsin Inhibitor	NKPLVVQFQK	10 fmol	$[M+2H]^{++} \rightarrow y_5^{+}$	100 amol	28,27
			$[M+2H]^{++} \rightarrow y_6^{+}$	100 amol	28,27
			$[M+2H]^{++} \rightarrow y_8^{+}$	10 amol	28,27
Trypsin Inhibitor	AAPTGNERCPLTWQSR	100 fmol	$[M+2H]^{++} \rightarrow y_8^{+}$	1 fmol	28,00
			$[M+2H]^{++} \rightarrow y_9^{+}$	1 fmol	28,00
			$[M+2H]^{++} \rightarrow y_{10}^{++}$	1 fmol	28,00

played clear isotopic clusters at 10 femtomoles and above. Similarly, one trypsin inhibitor peptide was visible at 1 femtomole of injected protein, the other two at 10 and 100 femtomoles. Thus, we concluded that relative quantification using an MS1 based label-free approach can be performed from amounts of 1 femtomoles and above. If the upper limit is arbitrarily set to 1 pmol, relative quantification could be done over a concentration range of three orders of magnitude (1 femtomole to 1 picomole).

Individual fragment ion currents were extracted for the same standard protein peptides from the *pseudo*-mSRM data. For three out of the six peptides, fragment ions were clearly visible at injected amounts of 10 atomoles and above (Table 1). The extracted fragment ion chromatograms for β-lactoglobulin peptide LFSNPTQLEEQCHI are shown in Figure 2.

Next, we investigated if the extracted fragment ion chromatograms could be used for label-free MS2-based relative quantification. However, this strategy is complicated by the fact that multiple fragment-ion chromatograms can be reconstructed from a single peptide precursor ion. Thus, we applied an empirical method similar to the one showed by Silva and co-workers. Indeed, this group showed that the average MS signal response of the three most abundant peptides is constant for each protein [9]. Based on this, the sum of the three most abundant fragment ions was calculated for all three peptides per standard protein. The protein abundance values calculated with this method showed excellent linearity with protein concentration (Figure 3).

Coefficients of correlation are above 0.98 for both measured proteins and all replicates over the full range of concentration. This implies that relative quantification using an MS2-based label-free approach can be performed at lowest protein concentrations. With an arbitrary upper limit of one

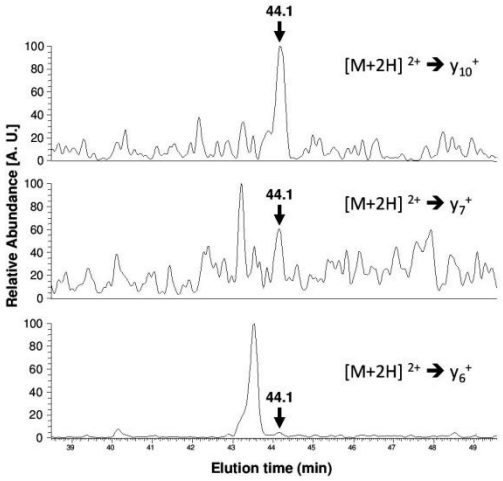


Figure 2. Extracted *p*-mSRM transitions ion currents from 10 amol of β-lactoglobulin peptide LFSNPTQLEEQCHI spiked into digested human plasma. The peptide elutes at 44.1 minutes.

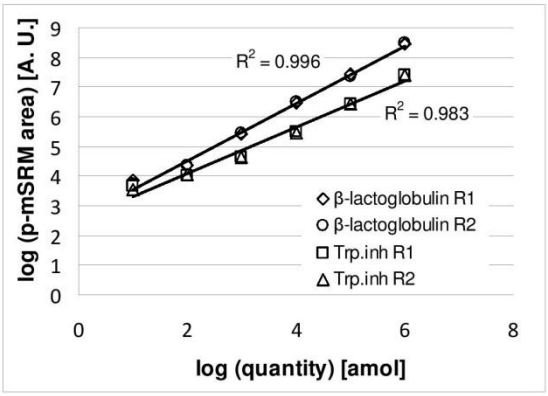


Figure 3. Combined *pseudo*-mSRM area as function of protein concentration for β-lactoglobulin and trypsin inhibitor spiked into human plasma. R1 = replicate 1, R2 = replicate 2. The correlation coefficient R^2 is calculated from the $\log_{(10)}$ values of *p*-mSRM area and concentration.

picomole, relative quantification could be done over five orders of magnitude (10 atomoles to 1 picomole). It represents an increase in dynamic range of two orders of magnitude compared to DDA acquisition, traditionally used for label-free quantification. Indeed, if the DDA mode is used on a very complex sample, the dynamic range of the analysis is close to the intra-spectrum dynamic range of the analyzer (in our case the Orbitrap analyzer). Makarov and co-workers have shown that the intra-spectrum dynamic range of an orbitrap was around 5,000 for full MS1 survey scans [18]. By contrast, the high dynamic range using the *pseudo*-mSRM mode can be explained by the combined dynamic range of the mass analyzer (in this case the linear ion-trap) and the ion injection time to fill the trap. The dynamic range of an ion trap is around $10^2 - 10^3$, with a varying injection time between 0.1-100 ms. Consequently, the resulting dynamic range with *pseudo*-mSRM is the product of this two values. This indicates that a dynamic range of five orders of magnitude should be possible, which corresponds to the value found in our experiments.

4. Conclusions

Our data shows that relative label-free quantification is possible in ion trapping devices, using the data-independent or *p*-mSRM mode of operation. Quantification can be performed directly, in complex samples over a large dynamic range. Moreover, combination of large-scale data-independent protein identification and label-free quantification is also possible. Data-independent acquisition strategies such as the recently published PAcIFIC provide good results in terms of number of identifications and dynamic range. The data format itself corresponds to nothing else than a large-scale *pseudo*-mSRM experiments. The obtained data can therefore be directly used for MS2-based relative quantification.

JIOMICS | VOL 1 | ISSUE 2 | DECEMBER 2011 | 211- 215

Acknowledgments

A. S. thanks the Swiss National Science Foundation for support (grant 315230_130830).

References

1. D.C. Stahl, K.M. Swiderek, M.T. DavisT.D. Lee, J. Americ. Soc. Mass Spectrom. 7 (1996) 532-540.
2. W. Wang, H. Zhou, H. Lin, S. Roy, T.A. Shaler, L.R. Hill, S. Norton, P. Kumar, M. Anderle, C.H. Becker, Anal Chem 75 (2003) 4818-4826.
3. D. Radulovic, S. Jelveh, S. Ryu, T.G. Hamilton, E. Foss, Y. MaoA. Emili, Mol Cell Proteomics 3 (2004) 984-997.
4. C.L. Gatlin, J.K. Eng, S.T. Cross, J.C. DetterJ.R. Yates, 3rd, Anal Chem 72 (2000) 757-763.
5. H. Liu, R.G. Sadygov, J.R. Yates, 3rd, Anal Chem 76 (2004) 4193-4201.
6. A. Panchaud, A. Scherl, S.A. Shaffer, P.D. von Haller, H.D. Kulasekara, S.I. Miller, D.R. Goodlett, Anal Chem 81 (2009) 6481-6488.
7. M. Bern, G. Finney, M.R. Hoopmann, G. Merrihew, M.J. Toth, M.J. MacCoss, Anal Chem 82 833-841.
8. S. Purvine, J.T. Eppel, E.C. Yi, D.R. Goodlett, Proteomics 3 (2003) 847-850.
9. J.C. Silva, M.V. Gorenstein, G.Z. Li, J.P. Vissers, S.J. Geromanos, Mol Cell Proteomics 5 (2006) 144-156.
10. A. Scherl, S.A. Shaffer, G.K. Taylor, H.D. Kulasekara, S.I. Miller, D.R. Goodlett, Anal Chem 80 (2008) 1182-1191.
11. V. Lange, P. Picotti, B. Domon, R. Aebersold, Mol Syst Biol 4 (2008) 222.
12. S. Yang, J. Cha, K. Carlson, Rapid Commun Mass Spectrom 18 (2004) 2131-2145.
13. J.H. Baek, H. Kim, B. Shin, M.H. Yu, J Proteome Res 8 (2009) 3625-3632.
14. P. Shipkova, D.M. Drexler, R. Langish, J. Smalley, M.E. SalyanM. Sanders, Rapid Commun Mass Spectrom 22 (2008) 1359 -1366.
15. L. Dayon, C. Pasquarello, C. Hoogland, J.C. Sanchez, A. Scherl, J Proteomics 73 (2010) 769-777.
16. A. Scherl, Y.S. Tsai, S.A. Shaffer, D.R. Goodlett, Proteomics 8 (2008) 2791-2797.
17. J.E. Elias, S.P. Gygi, Nat Methods 4 (2007) 207-214.
18. A. Makarov, E. Denisov, O. Lange, S. Horning, J Am Soc Mass Spectrom 17 (2006) 977-982.

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

In this chapter, we extended the application of pseudo-multiple selected reaction monitoring acquisition to large-scale bottom-up proteomics. We used the precursor ion independent from ion count (PAcIFIC) ion sampling method that uses a narrow isolation window not centered on the peptide *m/z* value, called ion channel. A list of incremental ion channels is used to cover a limited precursor ion *m/z* range per injection. Multiple injections of the sample are then performed to cover the full precursor ion *m/z* range. Due to the multiple injections of the same samples and the nature of data-independent acquisitions, the volume of acquired data is large. In addition, chimeric spectra are often acquired due to the nature of the ion sampling method. We thus developed a suit of algorithms to process data-independent tandem mass spectra.

This suit of algorithm consists of data clustering and precursor ion calculation from tandem mass spectra. The clustering method uses the grouping of co-eluting product ions to generate spectral networks of tandem mass spectra, prior to consensus spectrum building. The calculation of precursor ion *m/z* uses the complementarity of product ions from the cleavage of the same peptide bond from consensus tandem mass spectrum. We concluded that the application of this processing method reduces the data size, improves the number identifications and decreases the number of multiple hits per consensus spectrum compared to non-processed tandem mass data.

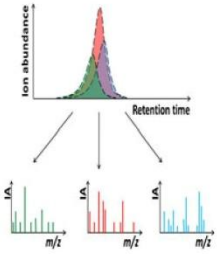
J. Am. Soc. Mass Spectrom. (2013) 24:1862–1871
DOI: 10.1007/s13361-013-0720-z

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]

[1]University of Geneva, Geneva, Switzerland
[2]SIB Swiss Institute of Bioinformatics, University Medical Center, 1, Rue Michel-Servet, 1211 Geneva 4, Switzerland
[3]Swiss 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@isb-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 10^7 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 µm, i.d. = 75 µm, L =150 mm) and pre-column (o.d. = 375 µm, i.d. = 100 µm, L =2 0 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 desalted 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 NanoAcquity chromatograph (Waters, Milford, MA, USA) interfaced with an LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific, San Jose, CA, USA). Peptides were trapped on a home-made, 20 mm long precolumn of 100 µm i.d. and separated on a 150 mm analytical column of 75 µm 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 ±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±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

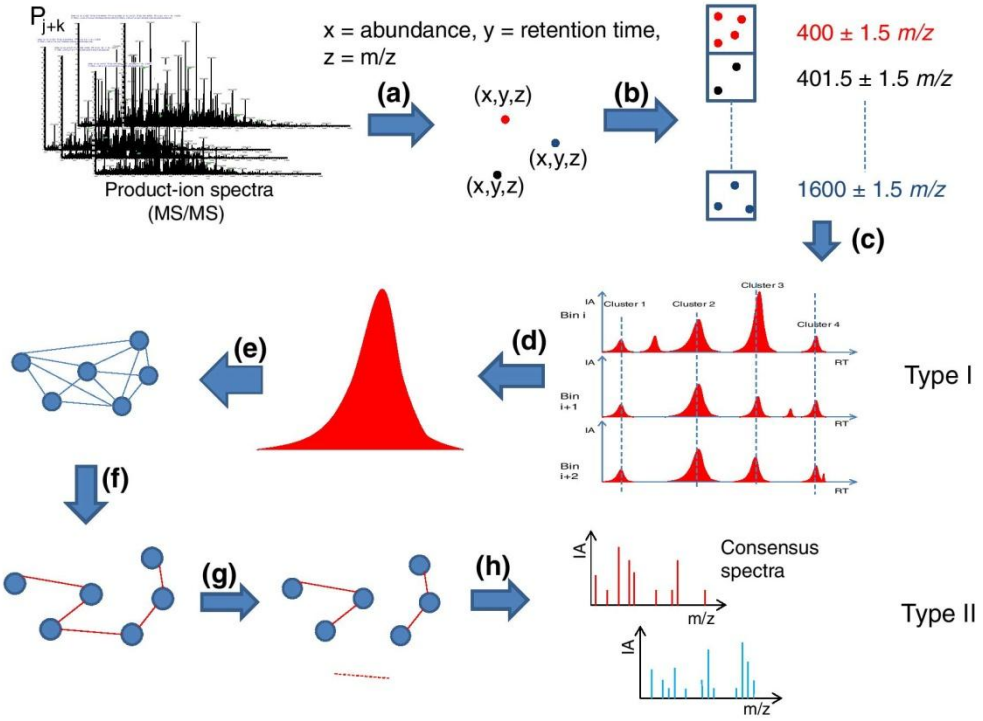


Figure 1. Tandem mass spectra are divided according to their precursor channels, then processed **(a)**. From a set of spectra, peaks are extracted for binning **(b)** and extracted FICs are generated for selecting local maxima **(c)**. An alignment of extracted FICs is performed to extract profiles 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 ($\Delta t = \pm 3$–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

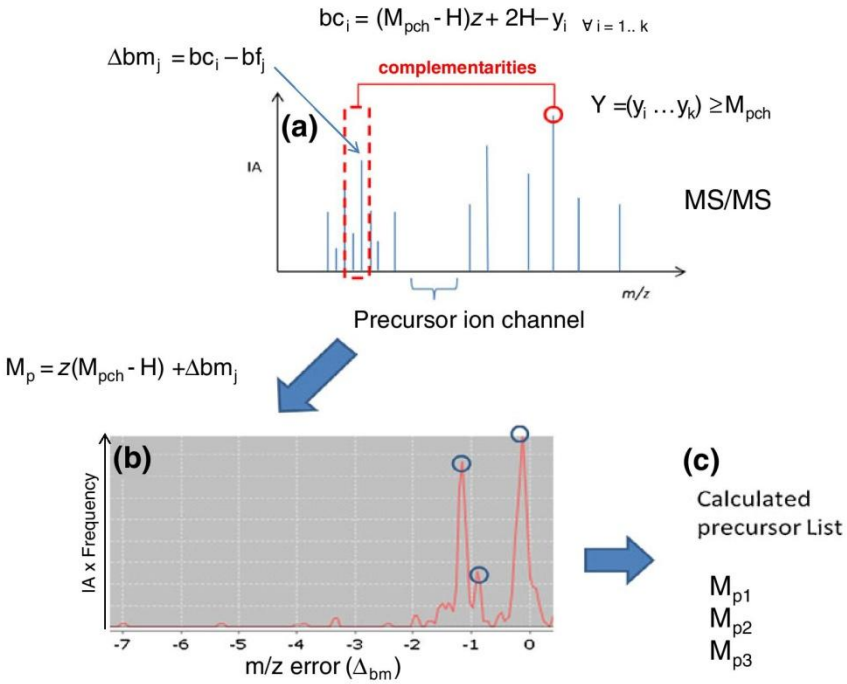


Figure 2. Overview of precursor ion correction. For a given spectrum, the precursor ion *m/z* value is estimated by the complementarities of b- and y-ions. **(a)** The center of precursor channel is used to calculate b-ion, and all possible shifts within a given window around this b-ion value are reported 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 ($Y = [y_i \ldots y_k]$), 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 ($bc_i \ldots bc_k$) 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 ($bc_i \ldots bc_k$) 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 ($\Delta bm_1, \ldots, \Delta bm_r$) between bc_i and the fragment *m/z* value ($bf_j, \ldots, bk_l$) together with the intensity of the fragments are stored (see Figure 2a).

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

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.http://javaprotlib.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 where 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

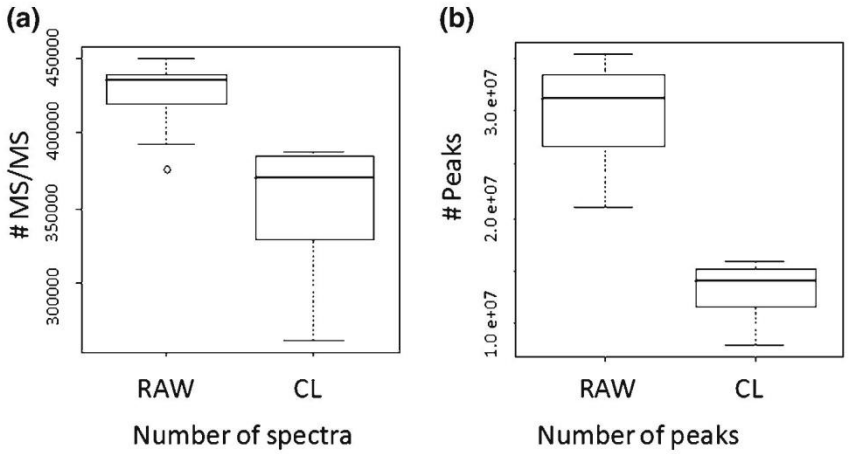


Figure 3. Data reduction after type II clustering (CL). The MS/MS spectra and peak counts are obtained from all 22 fractions. The boxes show the median, lower and upper quartiles of all 22 samples. Extreme values are indicated by horizontal lines and outliers by small circles. **(a)** The number of MS/MS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

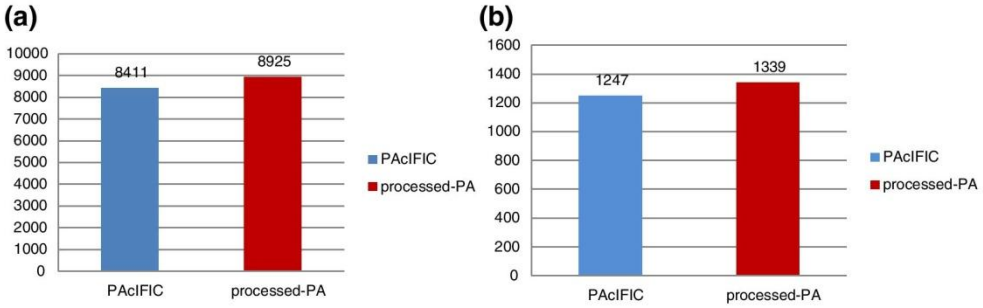


Figure 4. Unique peptide at a FDR of 5 % and protein identifications with at least two different peptide hits. Left **(a)**: unique peptide identifications. Right **(b)**: protein identifications. Processed PAcIFIC (processed-PA) data with type II data clustering shows highest number of identifications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

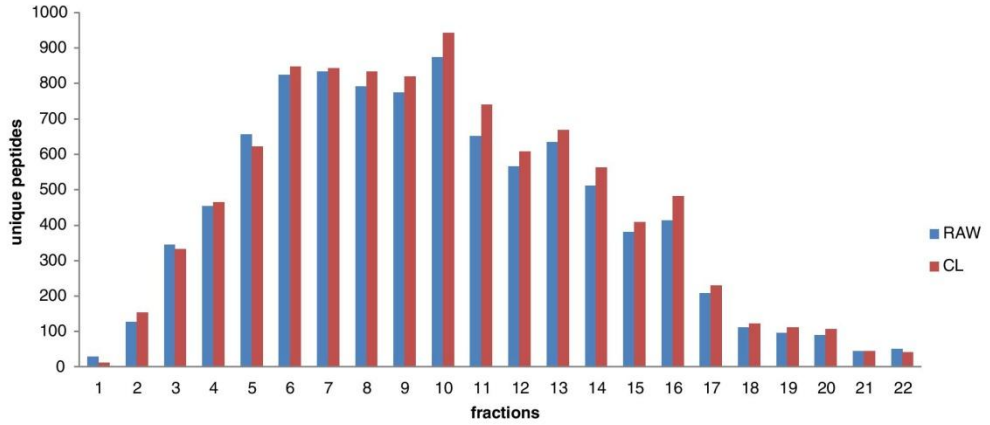


Figure 5. Distribution of the number of unique peptides identified in each fraction. The number of peptides is obtained at a false discovery rate <5 %. In most cases processed data show more identifications

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

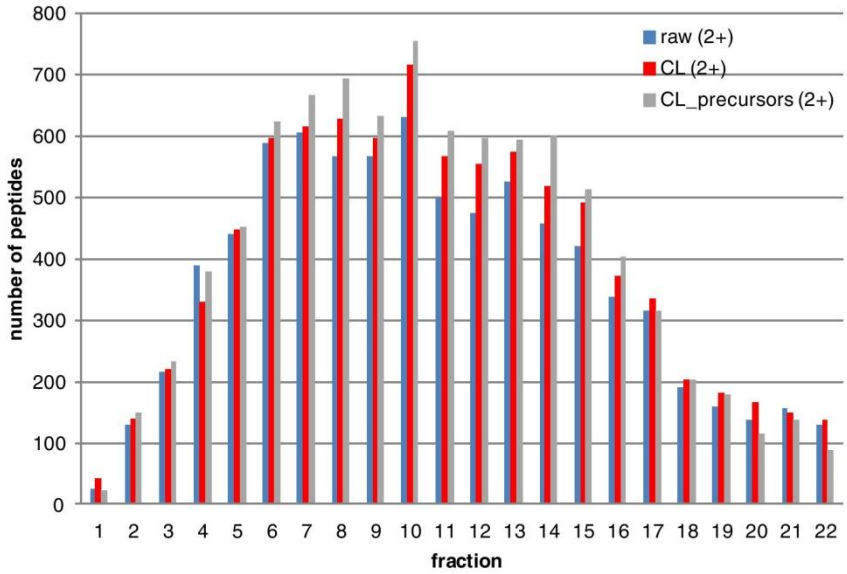


Figure 6. Distribution of unique peptides per fraction for doubly charged spectra. In blue, results from raw MS/MS spectra (raw [2+]). In red, results from MS/MS spectra after data clustering (type II) (CL [2+]). In grey, results after data clustering and precursor correction (CL_precursor [2+]). In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

1. 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)
2. 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)
3. 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)
4. 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)
5. 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)
6. 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)
7. 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)
8. 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)
9. 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)
10. 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)
11. 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)
12. 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)
13. 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)
14. 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)
15. 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)
16. 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)

Supplementary data

(http://link.springer.com/article/10.1007/s13361-013-0720-z)

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]**

[1]University of Geneva, Geneva, Switzerland

[2]SIB Swiss institute of Bioinformatics, Geneva, Switzerland

[3]Swiss Centre for Applied Human Toxicology, Switzerland

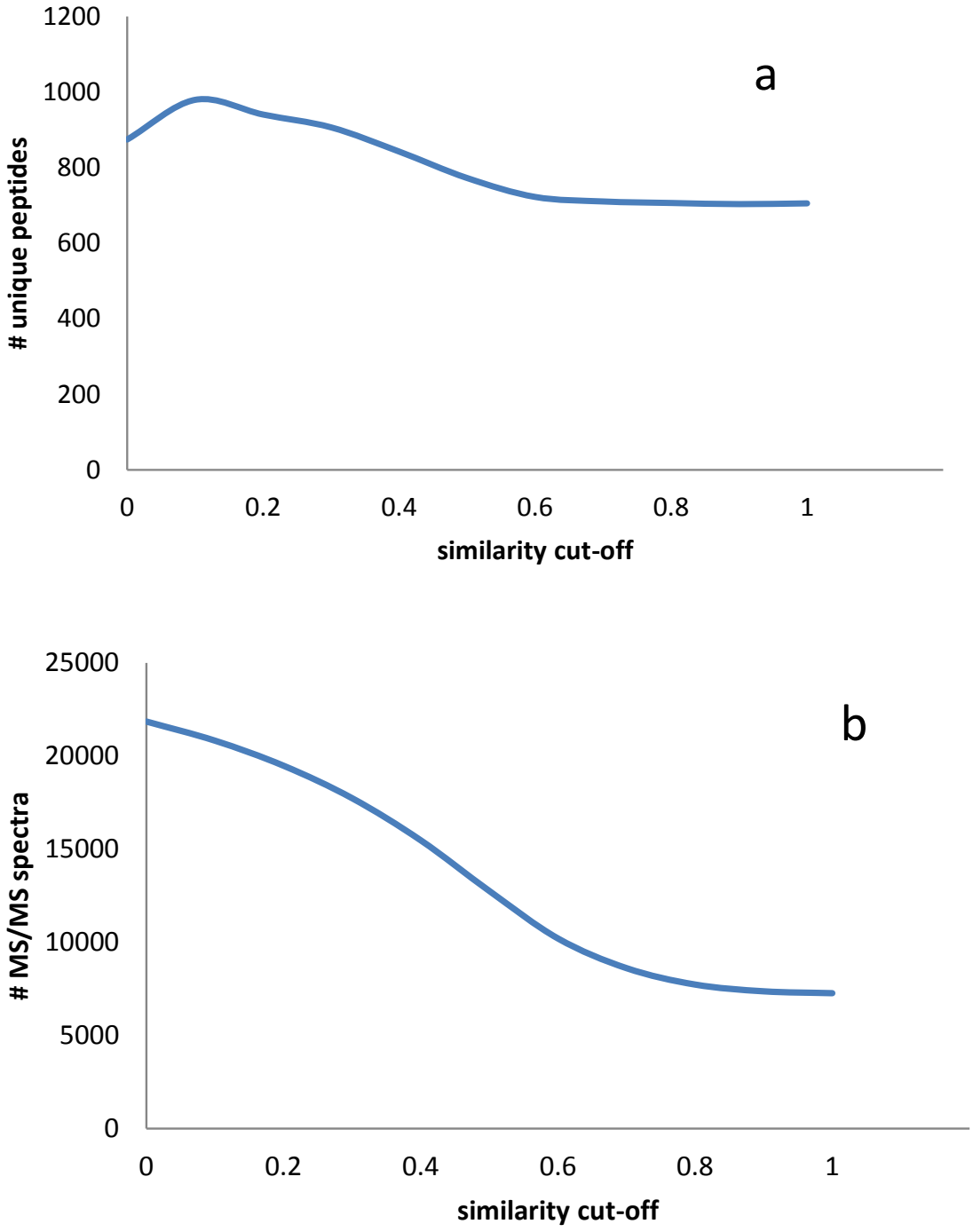


Figure S-1. Dependency of the number of MS/MS spectra and peptides identifications as a function of the similarity cut-off. A) Number of unique peptides identified at FDR = 5% f. B) Number of submitted MS/SM spectra.

$$\text{Spectrum } S^1 = \left\{P_i^1 = \left(m_i^1, I_i^1\right); i = 1 \cdots n_1\right\} \xrightarrow[binning]{} \vec{S}^1 = \left(s_1^1, s_2^1, .., s_N^1\right)$$

$$\text{Spectrum } S^2 = \left\{P_i^2 = \left(m_i^2, I_i^2\right); i = 1 \cdots n_2\right\} \xrightarrow[binning]{} \vec{S}^2 = \left(s_1^2, s_2^2, .., s_N^2\right)$$

$$s_j^k = \sum_{m_{\min} + j\Delta \le m_i^k < m_{\min} + (j+1)\Delta} I_i^k$$

$$score = \cos(\theta) = \frac{\vec{S}^1 \cdot \vec{S}^2}{\left|\vec{S}^1\right| \cdot \left|\vec{S}^2\right|} = \frac{\sum_{i=1}^{N} s_i^1 s_i^2}{\left(\sum_{i=1}^{N} s_i^1 s_i^1\right)^{1/2} \cdot \left(\sum_{i=1}^{N} s_i^2 s_i^2\right)^{1/2}}$$

Equation S-1. Normalized dot product score.

$S = s_i \ldots s_n$	# spectra in the file
M_{pch}	precursor channel of a given MS/MS spectrum
$Y = y_i \ldots y_k$	fragment y-ions with $y_i > M_{pch}$ of a given spectrum
$bc_i \ldots bc_k$	complementary fragment b-ions
b_window	window around bc_i
$bf_j \ldots bf_l$	b-ion fragments within a given window
$\Delta bm_j \ldots \Delta bm_l$	deviation from the center of bc_i
precursors	list of calculated precursors with associated Δbm and intensities
z	charge state (2+)

begin

for ($s_i \ldots s_n$)

for ($y_i \ldots y_k$)

$bc_i = (M_{pch} - H)z + 2H - y_i$

Find_b_peaks **($bc_i \pm$ b_window)** //find all peaks within the window

for ($bf_j \ldots bf_l$)

$\Delta bm_j = bc_i - bf_j$

$precursor_{calc} = M_{pch} + \Delta bm_j$ //new calculated precursor

$I_j = bf_j + y_i$ //intensity of $precursor_{calc}$

precursors.add(Δbm_i, $precursor_{calc}$, I_j)

Histogram (precursors) //plot 2D histogram precursors

Local_maxima (precursors) //select two precursors from histogram data

Write_spectra (s_i) //write MS/MS spectra with new calculated precursors

end

Figure S-2. Precursor ions *m/z* calculation is described here as pseudocode. Details are given in the materials and methods.

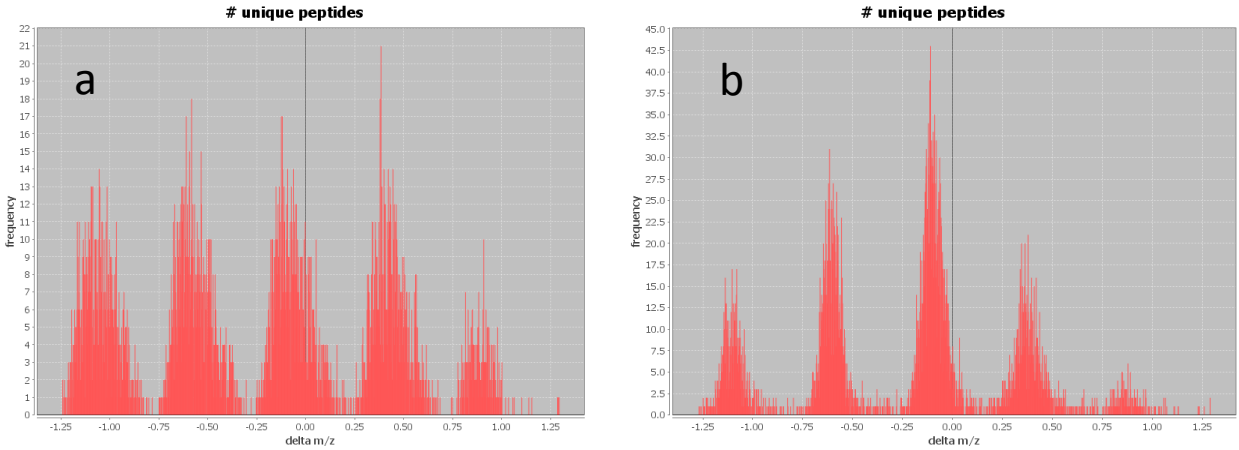


Figure S-3. Histogram of the differences between measured and theoretical precursor ion *m/z*. (A) Distribution without data clustering and precursor ion correction. (B) Distribution with data clustering and precursor ion correction. As shown in B), the distribution is more centered around 0 after data clustering and precursor ion correction.

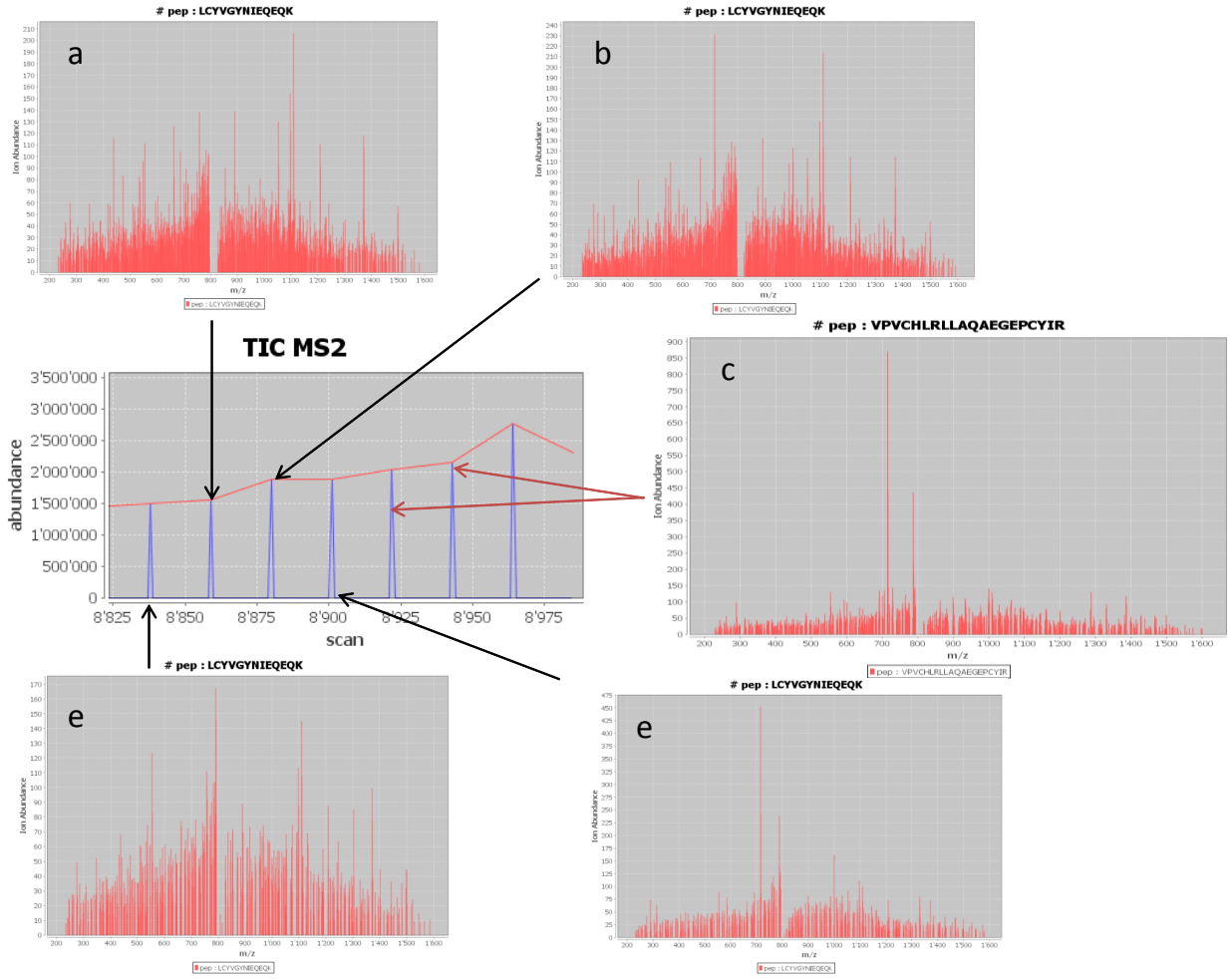


Figure S-4. A time window of precursor ion chromatogram of 822 *m/z* (TIC MS2). All spectra from scan 8800 to 8950 (A - E) are identified by type II spectra while only 2 spectra (A and B) are identified by non-processed spectra. Additionally 1 type II spectrum is formed from 2 spectra and matches 1 unique peptide specific to type II (C).

	up	g1	up_g1	r_g1	g2	up_g2	r_g2	g3	up_g3	r_g3
F9	29	55	23	32	5	3	7	1	2	1
F9_cl	13	18	10	8	2	2	2	0	0	0
F10	128	263	119	144	10	7	13	1	1	2
F10_cl	155	251	146	105	10	8	12	0	0	0
F11	346	797	309	488	64	45	83	15	13	32
F11_cl	333	566	305	261	36	27	45	4	6	6
F12	454	1054	424	630	63	39	87	4	9	3
F12_cl	465	883	442	441	38	26	50	2	6	0
F13	657	1735	614	1121	88	84	92	13	9	30
F13_cl	623	1251	575	676	57	76	38	8	10	14
F14	824	2347	763	1584	116	115	117	22	66	60
F14_cl	848	1825	794	1031	94	84	104	6	4	14
F15	834	2263	769	1494	123	92	154	19	27	30
F15_cl	843	1807	790	1017	105	82	128	10	14	16
F16	792	2198	748	1450	83	106	60	8	8	16
F16_cl	834	1865	793	1072	63	72	54	2	5	1
F17	775	2097	716	1381	171	121	221	16	14	34
F17_cl	820	1773	761	1012	127	87	167	11	11	22
F18	875	2590	830	1760	117	110	124	4	10	2
F18_cl	943	2242	889	1353	92	109	75	3	8	1
F19	652	2505	617	1888	88	65	111	4	10	2
F19_cl	740	1836	703	1133	361	68	654	22	13	53
F20	566	1854	540	1314	67	66	68	8	2	22
F20_cl	609	1615	586	1029	47	44	50	6	2	16
F21	635	2849	608	2241	100	57	143	5	3	12
F21_cl	669	2523	637	1886	59	58	60	12	1	35
F22	512	1454	498	956	59	27	91	1	3	0
F22_cl	564	1300	549	751	33	30	36	1	3	0
F23	381	1184	374	810	15	14	16	2	4	2
F23_cl	409	1048	397	651	14	17	11	2	3	3
F24	414	1585	407	1178	68	15	121	2	3	3
F24_cl	482	1481	472	1009	82	17	147	3	6	3
F25	208	755	202	553	10	5	15	0	0	0
F25_cl	230	623	225	398	9	5	13	0	0	0
F26	113	492	110	382	3	2	4	0	0	0
F26_cl	123	370	120	250	3	4	2	0	0	0
F27	96	304	95	209	0	0	0	0	0	0
F27_cl	113	242	112	130	0	0	0	0	0	0
F28	91	462	90	372	0	0	0	0	0	0
F28_cl	108	435	105	330	1	2	0	0	0	0
F29	46	92	45	47	0	0	0	0	0	0
F29_cl	46	68	45	23	0	0	0	0	0	0
F30	52	118	49	69	2	2	2	0	0	0
F30_cl	42	81	41	40	0	0	0	0	0	0

Table S-1. Summary of unique peptide identifications per spectrum and fraction. up = unique peptides, g1 = nb of spectra with 1 identification per scan (group 1), up_g1 = nb unique peptides among group 1, r_g1 = nb redundant spectra among group 1, g2 = nb of spectra with 2 identifications per scan (group 2), up_g2 = nb of unique peptides in group 2, r_g2 = nb redundant spectra among group 2, g3 = nb of spectra with 3 identifications per scan (group 3), up_g3 = nb of unique peptides in group 3, r_g3 = nb redundant spectra among group 3.

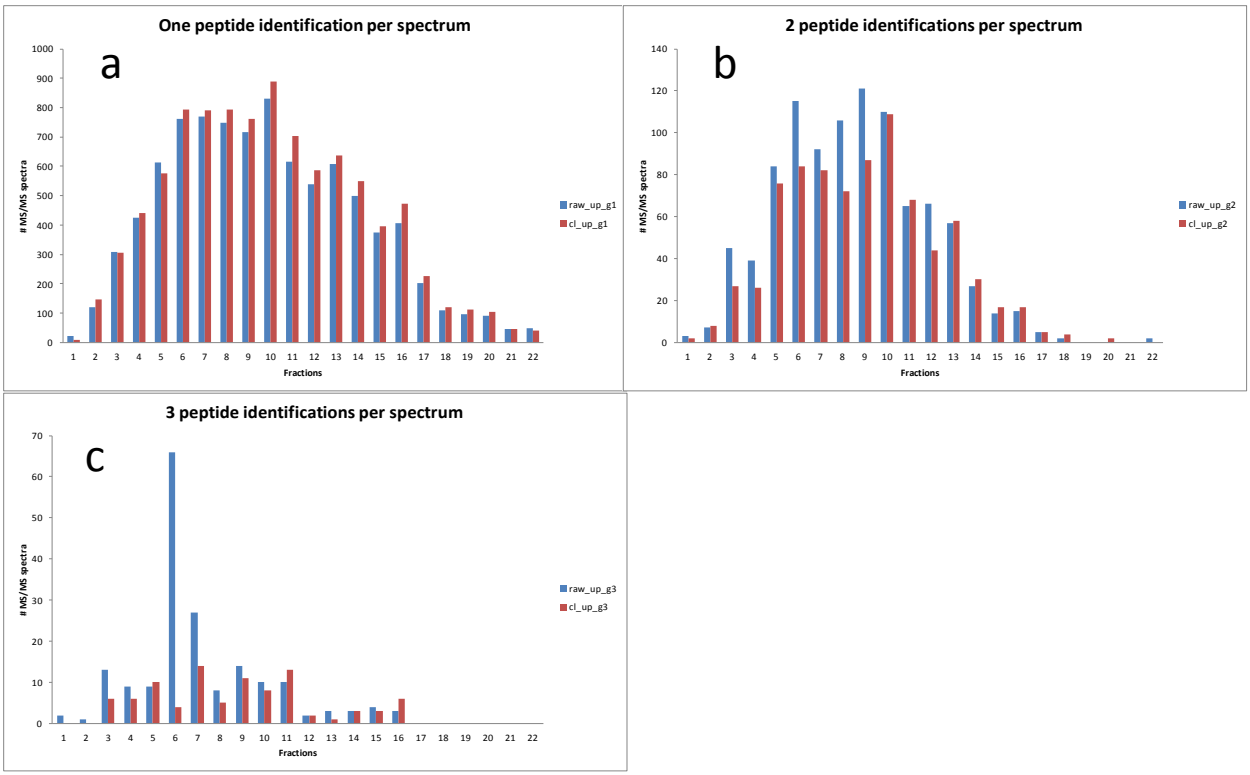


Figure S-5. Peptide identifications per spectrum. One peptide per spectrum (A), 2 peptides per spectrum (B) and 3 peptides per spectrum (C).

4 Protein modifications during the storage of red blood cells

In this chapter, we applied data-dependent and data-independent ion sampling methods to analyze red blood cells aging during blood storage. From a clinical point of view, the safety and the quality of red blood cells are important for transfusion. Several research papers have reported biochemical changes occurring in red blood cells that could disturb their functions and safety due to storage lesions.

For a comprehensive analysis of red blood cell proteins, we investigated the chemical modifications on them potentially related to prolonged storage. To detect possible modifications, we combined label-free quantification with open modification search of peptides. The integration of differentially abundant ion signals with open modification search revealed 1797 tandem mass spectra with mass-shifts ranging from -200 to 200 Da. For most of them, the nature of modifications was unknown. Nevertheless, for some of them, the modifications could correspond to ethylations, di-methylations, and other modifications related either to sample preparation or to storage lesion. We postulate that certain peptide modification could be potential biological markers of red blood cell aging.

Protein modifications during the storage of red Blood Cells

HuiSong Pak [1, 2], Pierre Lescuyer [1, 3], Sophie Waldvogel-Abramowsky [3], Markus Mueller [4], Alexander Scherl [1, 2, 3]

[1] Clinical proteomics and chemistry, Department of Human Protein Sciences, Faculty of Medicine, Geneva University, Geneva, Switzerland;

[2] Swiss Centers for Human Applied Toxicology, Geneva, Switzerland.

[3] Geneva University Hospitals, Geneva, Switzerland;

[4] Swiss Institutes of Bioinformatics, Geneva, Switzerland

Corresponding author Alexander Scherl

Department of human protein sciences

Geneva university hospitals

Geneva, university of Geneva

Rue Michel-Servet 1

CH - 1211 Genève 4

Phone # +41 22 372 73 61

E-mail address: Alexander.Scherl@unige.ch

Abbreviation:

RBCs: red blood cells; **DDA**: data-dependent acquisition; **DIA**: data-independent acquisition;

PAcIFIC: precursor acquisition independent from ion count; **MS**: mass spectrometry;

MS/MS: tandem mass spectrometry; **LC**: liquid chromatography

Keywords:

red blood cells, data-independent acquisition, data-dependent acquisition, open modification search, label-free quantification

Abstract

Prolonged duration of red blood cells was associated with the increase of molecular changes in the erythrocyte cells and the storage medium. Label-free tandem mass spectrometry analysis was used to screen for chemical modification occurring during the storage of red blood cells. In order to maximize identifications of possible modifications, samples were both analyzed in data-dependent and data-independent acquisition mode. The resulting tandem mass spectra were submitted to database search, and a spectral library was built from the resulting identifications. An open modification search was then performed to identify unexpected peptide modifications. The data was then integrated with quantitative information derived from label-free relative quantification between fresh and old red blood cell bags. With this strategy, we identified peptides and proteins present in different amounts and detected chemical modifications potentially occurring during the storage. Proteins involved in cellular oxidative stress and structural changes of the red blood cell membrane are present at higher amount after prolonged storage. The proportion of known and/or unexpected peptide modifications increased from 18% to 41% during this time. The nature of potential modifications could correspond to methylations, dimethylations, artefacts (*e.g.* sample preparation) and unreported modifications. In particular, we report the presence of peptides from protein PTIM1 at different abundance during the storage. Polymorphisme related to this protein was detected as modification and its abundance was measured. We postulate that identified and or/unknown peptide modifications might be dependent of red blood cell aging process during the storage lesion.

Introduction

Blood transfusion is important to save the life of critically ill patients in hospitals. For example patients in intensive care units are prone to be anemic and red blood cells (RBCs) transfusion is necessary to restore or increase tissues oxygenation [165][166]. In order to improve patients status, clinical studies were conducted on the maximal duration of RBCs in blood storage bags (age of RBCs). Aging of RBCs was associated with an increased risk of morbidity and/or mortality [167][168] for critically ill patients. A potential harmful effect of transfusing older blood due to RBCs storage lesions was shown. During the storage, RBCs are in a closed environment and suffer metabolic, biochemical and molecular changes that can disturb their functions. An increase of protein and lipid oxidation, as well as the presence of other biochemical substances have been reported in the storage medium [169][170]. In addition, RBC shape changes, membrane vesiculation and decreased deformability have also been observed [171]. Despite these reports that showed a deleterious effect of prolonging the duration of RBCs storage [172][173][174], many other studies demonstrated no effect of increasing the duration of RBCs storage [175][176][177]. The current debate about RBCs storage in blood transfusion medicine is how to avoid wasting RBC units and optimize the provision of blood stocks in hospitals. The standard practice consists of transfusing into patients oldest blood storage bags and storing RBCs units up to 42 days [178]. To answer to the question of safety and quality of long term stored RBCs, proteomics studies were conducted first at qualitative level to describe the diversity of proteins in the RBC proteome [179], then by addressing the issue of relative abundance at the protein level between fresh and old RBCs. For example, more than 350 membrane [180] and 1500 cytosolic [181] proteins were identified by liquid chromatography hyphenated with tandem mass spectrometry analysis (LC-MS/MS), showing the diversity and the complexity of the RBCs proteome. As reported by Roux-Dalvai and co-workers, the depletion of abundant proteins

such as hemoglobin (98 % w/w), showed an increase of identification of single gene products in the remaining 2 %. The dynamic of proteins in RBCs during the storage was measured and described by Zolla and co-workers, and Bosman and co-workers. The degradation of proteins such as spectrin and β-actin by oxidation has been reported, and storage-dependent changes of proteasome and chaperone proteins have also been observed [182][183]. All studies concluded to a modification of RBCs proteome during the storage due to cellular stress, indicating cellular aging process in terms of membrane or cytolsolic protein profiles or modifications.

From a clinical point of view, it is important to ensure the quality and safety of blood in blood banks before transfusing RBCs into patients. In this study, we attempted to look for a biological indicator of RBCs aging during the storage. In particular, we were interested in identifying possible chemical or biochemical modifications of RBC proteins rather than discovering previously undetected proteins (see **Figure 1**). For this, we combined conventional data-dependent acquisition (DDA) tandem mass spectrometry (MS/MS) and data-independent acquisition (DIA) tandem mass spectrometry analysis to compare fresh and old blood samples from blood banks. We used high-resolution mass spectrometry to analyze peptide profiles from trypsin-digested RBCs. Then, we measured peptide abundance profiles in both conditions, using label-free protein quantification methods. To identify as many modified peptides as possible potentially created during the storage, such as post-translational modification (PTM), chemical addition of products (adducts) and others, tandem mass spectrometry data were searched for unexpected modifications (open modification search, OMS). Information extracted from peptide abundance profiles and OMS were integrated to determine protein abundance profiles with expected and previously unknown modifications.

Materials and methods

Materials. Iodoacetamide (IA), acetonitrile, tris(2-chloroethyl) phosphate (TCEP), and hydroxylamine were purchased from Sigma (St. Louis, MO, USA). Urea, triethylammonium bicarbonate (TEAB), dithioerythriol (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 pre-columns were from Michrom (Auburn, CA, USA) and analytical column (o.d. = 375 µm, i.d. = 75 µm, l = 150 mm) was purchased from Nikkio (Japan). The pre-column (o.d. = 375 µm, i.d.= 100 µm, l = 20 mm) was made from fused silica tubing from BGB Analytik AG (Boeckten, Switzerland). Ultrasonicator was from Hieschler Ultrasound Technology (Teltow, Germany). Rapigest SF surfactant was from (Waters, Milford, MA, USA).

Samples. Fresh RBC bags (storage time < 5 days, n=3) and old RBC bags (storage time < 42 days, n=4) were obtained from the blood transfusion center of the Geneva University Hospital. Fresh samples were received as blood rods from different donors. Old samples were received as blood storage bags and the first blood rod from each bag was used for RBCs protein extraction. No ethical agreements were needed for the use of blood storage bags, because they were destined to be thrown away. We had no information regarding blood donors. For sample preparation, proteins were extracted according to previously published article from our group [184]. Briefly, 1000 uL of blood was centrifuged at 3500g for 10 min. Then plasma was removed. Red blood cells were suspended in 9000 uL of cold solution of 0.9% NaCL and centrifuged at 700g for 10 min. The supernatant was removed. This cleaning step was repeated two more times. Then RBCs were lysed by the addition of 8000 uL of cold

deionized water and centrifuged at 12’000 g for 10 min. All the operations of RBCs protein extractions and lyses were repeated seven times for three fresh and four old blood storage bags from Geneva University Hospitals blood bank. Soluble proteins were suspended in a surfactant containing ammonium bicarbonate buffer (Rapigest, Waters, Milford, USA, reference 186001860) (pH 8.0) and homogenized according to the manufacturer's instructions. Protein concentration was measured by Bradford quantification (BioRad, Hercules, USA, reference 500-0001). 100 µg of protein from each sample was then reduced using dichlorodiphenyltrichloroethane (50 mM) and alkylated by adding iodoacetamide (15 mM). Trypsin (1:50 w/w) was added and incubated overnight. Rapigest was then removed according to the manufacturer's protocol. Finally, samples were desalted using a C18 microspin column (Harvard, Holliston, USA, reference 747206) according to manufacturer's instructions. The dried material was then suspended in CH3CN/FA 2% / 0.1% prior to LC-MS/MS analysis.

Liquid chromatography andmass spectrometry. The liquid chromatography-tandem mass spectrometry (LC-MS/MS) system consists of a NanoAcquity chromatograph (Waters, Milford, MA, USA) interfaced with an LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific, San Jose, CA, USA). Peptides from pooled samples from fresh (n=3) and old (n=4) RBC bags were analyzed in triplicate. For this, they were trapped on a home-made, 20 mm long pre-column of 100 um i.d. and separated on a 150 mm analytical column of 75um inner diameter. The analytical separation was performed over 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 0-1min 95% A and 5% B, then to 65% A and 35% B at 55min, and 20% A and 80% B at 65 min at a flow rate of 220 nL/min. For data-dependent acquisition (DDA), a simple scan from 400-2000 *m/z* range was acquired in the Orbitrap, prior to 8 top abundant precursor ion

selections for product ion scans (MS^2) in the linear ion trap (LTQ). In total six replicates were analyzed by DDA, three injections for fresh samples and three injections for old samples. For DIA tandem mass spectrometry, we used the precursor acquisition independent from ion count (PAcIFIC) method [15]. Simple scans 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^2 spectra were acquired over 20 precursor ion channels in the linear ion trap for each LC-MS/MS analysis, with an isolation window of 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 ion mass range of 450-1100 *m/z* and injected the same sample 13 times. For example, the first injection (fraction 1) of our PAcIFIC [15] data set covers precursor *m/z* from 450-497.5 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 500 to 547.5 units. The same increments were used in the next fractions, until *m/z* = 1100 was reached.

Label-free quantitative LC-MS/MS data processing. Label-free LC-MS quantification was processed by Progenesis QI for proteomics (Waters, Milford, MA, USA) [185] with default parameters for high-resolution and high-mass accuracy data with a peak detection window larger than 0.15 min for MS^1 feature detections and quantifications. The software for alignment automatically chose the best ion chromatogram. One-way Analysis of variances (ANOVA) and principle component analysis (PCA) were applied on 13000 MS features for validation and separation of features respectively for old and fresh conditions. All MS features with p-value ≤ 0.05 and peptide fold-changes ≥ 5 were extracted (more than 3000 features).

Protein identifications. Peak lists were generated from raw data using the EasyProt protein identification platform [186]. The peaklist files were searched against the uniprot_sprot database (2013_12 of 11-Dec-2013). Homo sapiens taxonomy was specified for database searching. The parent ion tolerance was set to 10 ppm. Variable amino acid modifications were oxidized methionine. Trypsin was selected as the enzyme, with one potential missed cleavage, and the normal cleavage mode was used. The peptide FDR ratio was set to 1%. For all analyzes, only proteins matching two different peptide sequences were kept. For DIA MS^2 data, the same settings as for DDA data identification were used, except for precursor ion tolerance and false discovery rate (FDR). Those values were set to 1.3Da and 5%, respectively for mass tolerance and FDR.

Open Modification Search (OMS). Peptides were searched for unexpected modifications using the QuickMod software [187]. A sample specific spectral library was created with deliberator [188], using identifications from DDA and DIA acquisitions from protein pepXML files. Unmatched tandem mass spectrometry data was extracted from the database search, and then used for (OMS). A mass-shift of ± 200 Da was set for OMS with a mass accuracy of 50ppm. The integration of MS features and OMS data was based on the matching of precursor ion mass and retention time within a given tolerance of 10ppm and 1.5 min retention time for DDA acquisition.

Results

Identification of proteins by tandem mass spectrometry. Fist, the number of unique peptides and corresponding proteins were summed for each condition and for each data acquisition method. In total, 206 unique proteins were identified from fresh RBC bags and

174 proteins from old bags. Both acquired data sets resulted with lower numbers of peptide and protein identifications in old samples compared to fresh samples (See **Table 1)**. This might be related to the effect of storage lesion, leading to protein degradations or biochemical modifications. For DDA, the number of unique peptides and proteins are summed from six LC-MS/MS analyzes (three injections per condition). It resulted with the identification of 506 unique peptides and 132 proteins. For DIA, the number of unique peptides and proteins are summed from 26 LC-MS/MS analyzes (13 injections per condition). It resulted with 831 unique peptide identifications and 210 protein identifications. Out of these proteins, 126 (406 unique peptides) are found to be in common, six (100 unique peptides) unique to DDA and 84 (425 unique peptides) unique to DIA (see **Figure 2**). This high number of identifications from DIA is related to the systematic interrogation of ion mass channels, allowing detection of low abundance peptides. This is in agreement with other reports on data-independent acquisition methods [149]. In particular, PAcIFIC was reported to outperform other DIA methods because it uses narrow isolation windows for parallel activation of ions and allows also to submit tandem mass data to conventional sequence database search engines for peptide / protein identification [15]. In average, the unique peptide-to-protein ratio shows the redundant selection of abundant precursor ions and the low sampling power of DDA compared to DIA. Indeed 100 unique peptides identify six proteins in DDA, whereas 425 unique peptides identify 84 proteins (five unique peptides / protein). These results demonstrate the enhanced ions sampling method of DIA to analyze complex mixture of peptides.

Label-free quantification. LC-MS feature extractions and comparisons were performed on pooled fresh (n=3) and pooled old (n=4) blood samples. In total, 12899 features were detected from simple scans and used for relative quantification. Among all features, 4899 were identified with DDA and 8000 features had no corresponding identification. After ANOVA analysis to select MS features with a p-value ≤ 0.05 and fold-changes ≥ 5, 3028 MS

features were detected and extracted for tentative identification. We included this step to reduce the data and to keep only relevant features for the discovery of biological indicator of RBC aging (see **Figure 3**). In order to verify the separation of the 3028 features between old and fresh RBCs, principal component analysis was performed (see **Figure 4**). We could observe that the separation according to the first component explains 89 % of the variances, and 4 % of variances are explained by the second component. A list of 18 proteotypic peptides, corresponding to 13 proteins was extracted (see **Table 2**). After integration of all label-free LC-MS data with peptide identifications from database search, only seven peptides were found with higher abundance profile in old RBC samples. To verify the validity of our integrated data, we compared the extracted ion chromatograms of peptides of B3AT and 1433E proteins in old and fresh samples (see **Figure 5**). The measured peptide abundance ratio was different in absolute value compared to the automatically calculated value, but showed the same trend, an important increase in their peptide abundance in old blood storage bags.

Open modification search. For open modification search of peptides, unmatched tandem mass spectra from DDA after database search were extracted. Modifications of up to 200 Da were searched. The difference between fresh and old samples for open modification search is shown as mass-shift histograms (see **Figure 6**). Various numbers of mass-shifts, corresponding to known and/or unknown modifications are shown. In most cases, the mass-shifts are identical between old and fresh RBCs. The frequency of mass-shifts is higher for old RBCs compared to fresh RBCs, showing more frequent modifications. The resulting data was then used to match the 3028 MS features with significantly different abundance with peptide modifications identified by OMS. In total, 1797 features corresponding to unmatched peptides after database search with significantly different abundance between old and fresh samples were matched by OMS. After this step, modifications corresponding to known and

expected artifacts such as carbamylation, iodination, methionine oxydations as well as peptides from proteins never observed in RBCs were discarded. Next, the value of the mass shift was used for the hypothetical assignment of the nature of the modification. A summary of the peptides and nature of the modifications is shown in **Table 3.** While some mass shifts corresponded to methylation, dimethylations and ethylation, others did not correspond to known modifications. Finally, one mass shift of 14.01 Da could correspond either to a methylation of a serine residue or a substitution of a valine residue to isoleucine. These two modifications are isobaric. **Figure 7** shows the corresponding tandem mass spectrum, annotated with the two possible modifications.

Next, we verified the presence and the higher abundance of modified peptides from Table 3 in the pooled samples by comparing extracted ion chromatograms of the corresponding peptides. From the pooled samples, higher abundance of all modified peptide signals could be confirmed (**Table 3**) (see **Figure 8**).

Discussion.

In order to report differentially expressed peptides and/or their chemical modifications, protein digest from fresh and old RBCs were analyzed by label-free LC-MS/MS. The complementarity of DDA and DIA methods was used to increase the number of peptide identifications and consensus spectra for sample spectral library. Chemical modifications were detected by open modification search. The nature of the modification was tentatively determined using the value of the mass shift. The open modification search and label-free quantification data were integrated based on precursor ion mass accuracy, chromatographic retention time, minimum feature fold-change and dot product score threshold. In total, 12’899 MS1 features were matched to 14’433 consensus tandem mass spectra. It resulted in 555

unique matches from fresh samples and 1242 matches from old samples. This means that 555 peptide modifications were identified in pooled fresh samples and 1242 from pooled old samples. It represents respectively 3 % and 8 % of all consensus spectra from the sample spectral library. Among all modified peptides with different abundance between old and fresh samples, 18% were found in fresh RBCs and 41% in old RBCs. Thus, as suspected, the rate of chemical modifications increased with prolonged duration of blood storage bags. The average abundance of modified peptides is doubled in old storage bags and represents almost half of the abundant features. This could be an indication of biochemical changes due to storage lesion.

Only 19 non-modified peptides matched high abundant features. Among them, 17 peptides corresponding to 13 RBCs proteins were already described as being part of the RBC proteome [180][181]. All of them were involved in protein degradations, apoptosis, protection from oxidative process, oxido-reduction and membrane shape changes. For example, Band 3 protein is used as substrate by caspase 3, an important protein for cellular apoptosis. The protein Band 3 protein is also involved in membrane organization. Other examples are CAND1, which is a mediator of ubiquitination of proteins, and Peroxiredoxin-6, which has an anti-oxidative function. Regarding their abundance profile, some of them are probably false positive due to measurements from only one unique peptide and/or in contradiction with literature. For example, 1443E is considered as internal organelle protein and should probably not have been identified as being present with higher abundance in old RBC samples. Mature RBCs have no internal organelles and nucleus. However, RBCs are at different state of aging during the storage and should thus be considered as being in continuous development [180]. To have a correct protein expression profile, the quantification data could be cross-validated on the basis of label-free quantification and 2-DE gels [183][182]. Indeed, 2-DE gel

electrophoresis is a powerful tool to observe protein degradation and the behavior of multiple proteoforms.

We were not able to identify half of the remaining abundant features. We can indeed only speculate on the nature of possible reasons: chemical modifications of molecules from the attack of reactive oxygen species (ROS) and other oxidative process that leads to protein aggregation or degradations. These phenomena were previously observed on several proteins, such as Band 3, Band 4, spectrin and other cytosolic and membrane proteins. Band 3 proteins and spectrin were reported as main target of ROS attack and their degradation were confirmed by 2-DE gels of proteins [182]. As described by D'Amici and co-workers, a possible modification of the C-terminal domain of proteins by ROS attack was also reported. Indeed, these features could correspond to protein fragments after degradation of the C-terminal domain. Other modifications might origin from sample preparation artifacts, such as dicarbamidomethylation resulting with a mass-shift of 114.04 Da, isobaric with ubiquitinylation of K and S residue. Mass-shifts corresponding to iodination (125.89 Da) were also not considered because such modifications are probably not related to storage lesion. For the peptide ELVDD**S**VNNVR of protein L-isoaspartate (D-aspartate) methyltransferase, we can reasonably assume that the mass-shift of 14.01 Da corresponds to the substitution of V residue into I residue. The annotation of the tandem mass spectrum with this substitution result with a more abundant y_5 fragment ion compared to an adduct of 14.01 Da on the S residue. It was not possible to observe the presence of this peptide modification in fresh replicates. It was reported that the V variant has reduced activity and stability at varying temperature and in the presence of oxidative stress (*e.g.* H_2O_2) compared to I variant [189]. We can thus speculate that in the presence of oxidative process in RBCs, such as ROS attacks on proteins, the variant V was more subject to degradation than I variant. The protein L-isoasparte methyltransferase is involved in the repair and/or degradation of damaged proteins

in RBCs. The ratio of the variants could thus potentially be used as an indicator of RBC aging.

Nevertheless, several hundreds of mass-shifts are visible in the mass-shift histogram from open modification searches. For most of them, we currently have no information about their nature. Directed product ion scans with high measured mass accuracy and resolution could be used in the future to attempt identification of these unknown modifications.

Conclusion. In this study, the goal was to identify chemical modifications of RBC proteins during their storage in blood bank conditions. In order to achieve this goal, we acquired comprehensive mass spectrometry-based proteomics data using various ion sampling methods. In particular, we used data-independent acquisitions to achieve good proteome coverage while minimizing liquid handling and fractionation of the samples. Then, open modification searches were performed in addition to database searches in order to identify modified peptides. The combination of all these strategies allowed us to identify peptides carrying biochemical modifications probably related to storage lesion of RBC and potential markers of RBC aging. Compared to other studies using extensive fractionation and/or depletion of abundant proteins, the total number of identified proteins was relatively low. However, because sample preparation and liquid handling was kept to a minimum, we postulate that the identified modified peptides might be mostly specific to storage lesions and RBC aging.

List of tables:

	protein	merge	unique peptide	merge
RBC_DDA_OLD	96		385	
RBC_DDA_FRESH	121	132	461	506
RBC_DIA_FRESH	199		793	
RBC_DIA_OLD	169	210	641	831

Table 1: Protein and peptide identifications by DDA and DIA. The number of unique peptides and proteins for both acquisitions modes and both conditions are reported in the table.

peptide	proteotypique	proteins	protein name	location
EALQDVEDENQ	true	1433E_HUMAN	14-3-3 protein epsilon	membrane
TYDATTHFETTC[Cys_CAM]DDIK	true	GDIB_HUMAN	Rab GDP dissociation inhibitor beta	cytosol
LEAEGVPEVSEK	true	GLRX3_HUMAN	Glutaredoxin-3 (Glutaredoxin)	cytosol
SNLENIDFK	true	PUR8_HUMAN	Adenylosuccinate lyase	membrane
ASTPGAAAQIQEVK	true	B3AT_HUMAN	Band 3 anion transport protein	membrane
LQEAAELEAVELPVPIR	true		Band 3 anion transport protein	membrane
ADFLEQPVLGFVR	true		Band 3 anion transport protein	membrane
LYSNAYLNDLAGC[Cys_CAM]IK	true	AL1A1_HUMAN	Retinal dehydrogenase 1	cytosol
ITSEALLVTQQLVK	true	CAND1_HUMAN	Cullin-associated NEDD8-dissociated protein 1	cytosol
C[Cys_CAM]GEDDETIPSEYR	true	ICAL_HUMAN	Calpastatin	cytosol
IIFVVGGPGSGK	true	KAD1_HUMAN	Adenylate kinase isoenzyme 1	cytosol
IGQPTLLLYVDAGPETMTQR	true		Adenylate kinase isoenzyme 1	cytosol
DINAYNC[Cys_CAM]EEPTEK	true	PRDX6_HUMAN	Peroxiredoxin-6	cyotosol
QLC[Cys_CAM]DNAGFDATNILNK	true	TCPH_HUMAN	T-complex protein 1 subunit eta	cytosol
QQLLIGAYAK	true		T-complex protein 1 subunit eta	cytosol
GWEEGVAQMSVGQR	true	FKB1A_HUMAN	Peptidyl-prolyl cis-trans isomerase	membrane
PPYTVVYFPVR	true	GSTP1_HUMAN	Glutathione S-transferase P	cytosol

Table 2: Differentially measured abundant peptides in RBCs. The table summarizes peptide detected with different abundance in fresh (blue) and old (yellow) samples with one-way ANOVA p-value < 0.05 and peptide fold-change > 5. The names and subcellular locations of differentially abundant proteins are also reported in the table.

peptide	protein	modification	predicted	exp.mass	Δm (ppm)	mean fresh	mean old	ratio
G({27.99})GPFSDSYR	CAH1	NA	NA	507.2198	NA	2.58E+05	2.63E+06	10.19
E({28.00})SISVSSEQLAQFR	CAH1	ethylation?	804.8972	804.8958	1.74	4.50E+04	1.54E+06	34.3
H({28.00})DTSLKPISVSYNPATAK	CAH1	ethylation?	653.0081	653.0057	3.68	2.80E+05	3.54E+06	12.63
F({27.99})NTANDDNVTQVR	CATA	NA	NA	761.3474	NA	9.02E+04	2.92E+05	45.24
ELVDDSV({14.01})NNVR	PIMT	methylation / I-120	637.3212	637.3229	-2.67	3.70E+04	1.23E+06	33.13

Table 4: List of abundant peptide modifications in pooled fresh and old samples. The AUC of each peptide was measured by the extraction of extracted ion chromatogram from high-resolution simple scans.

List of Figures:

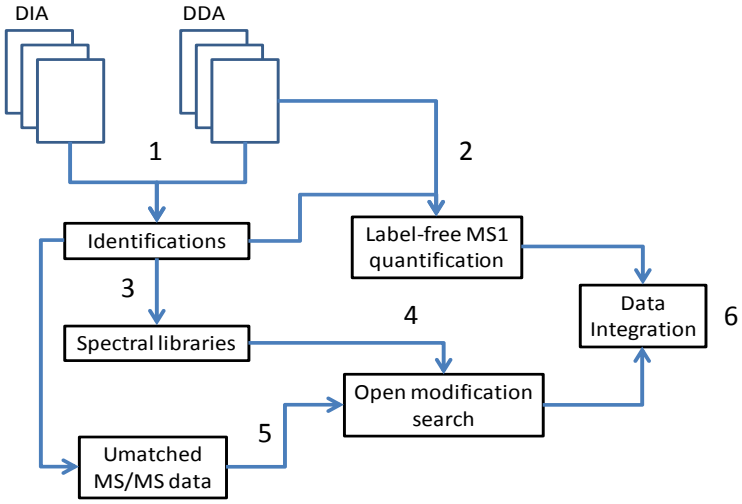


Figure 1: Data-dependent and data-independent acquisition strategy for the identification of RBC peptides. 1) MS^2 data from DDA and DIA are searched in sequence database for peptide identification. 2) In parallel, DDA data are processed for label-free peptide quantification. 3) Identifications from database searches are used to generate a sample-specific spectral library. 4) The spectral library is used for open modification searches (OMS). 5) Only unmatched spectra are used for OMS. 6) Quantitative data and identified modified peptides are integrated.

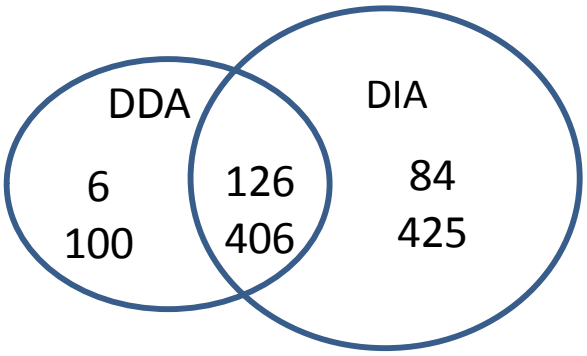


Figure 2: Venn diagram of all identified peptides by DDA and DIA tandem mass spectrometry. Number of unique peptides (numbers at the bottom) and proteins (numbers at the top) are reported in the diagram.

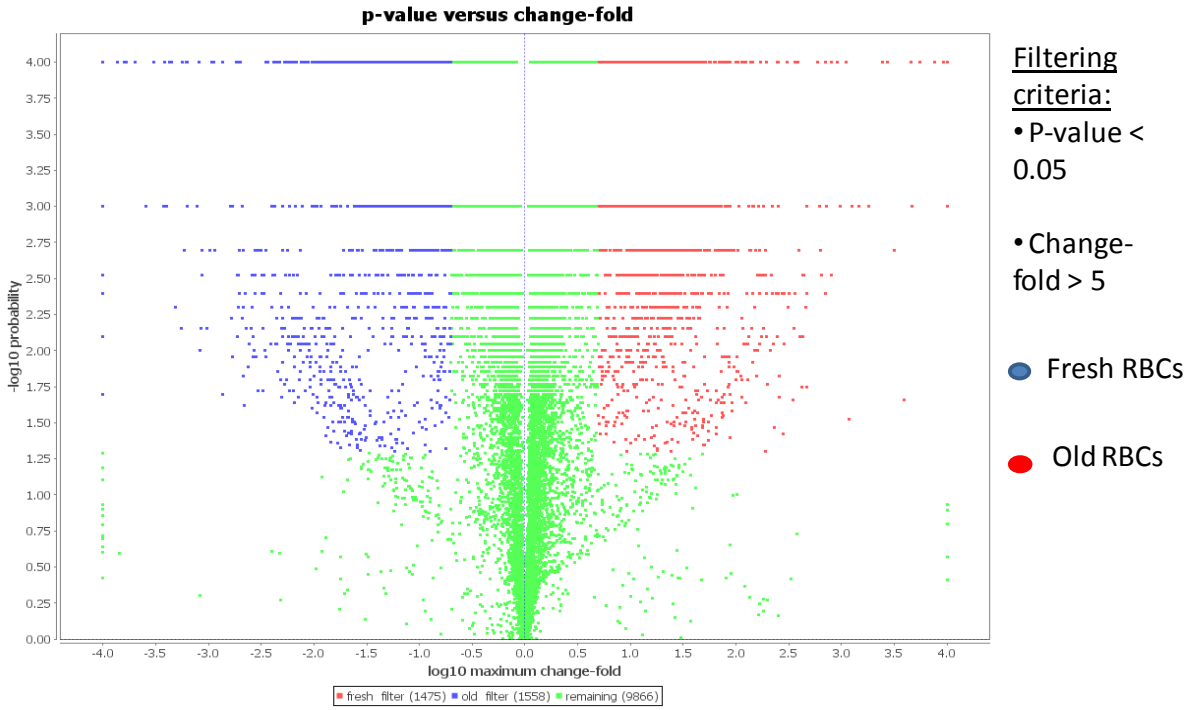


Figure 3: RBCs peptides vulcano plot. More than 13000 features (dots) were detected from blood samples. MS1 features with one-way ANOVA p-value < 0.05 and peptide fold-change > 5 (blue and red dots, respectively for fresh and old blood samples) are extracted (3028 MS features) prior to integration with OMS data.

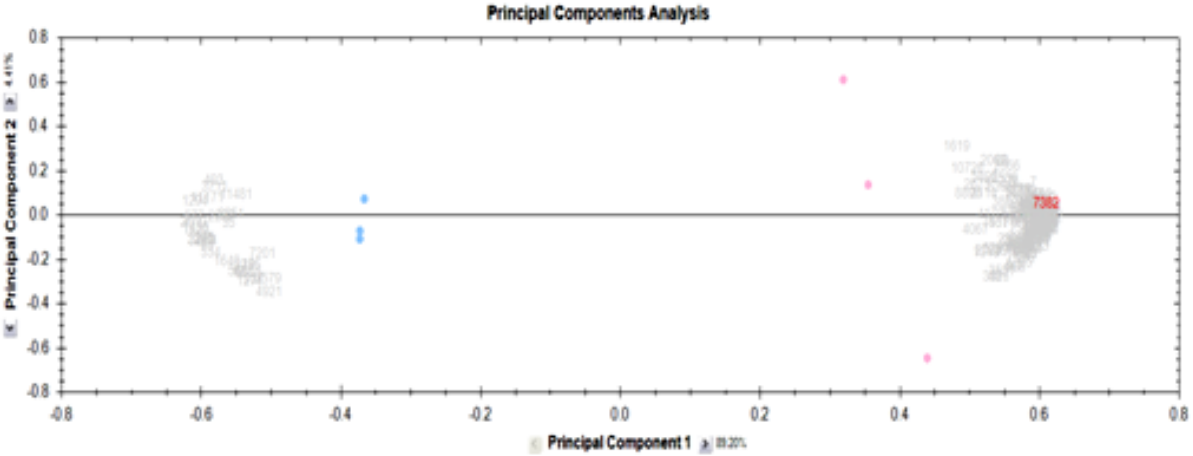
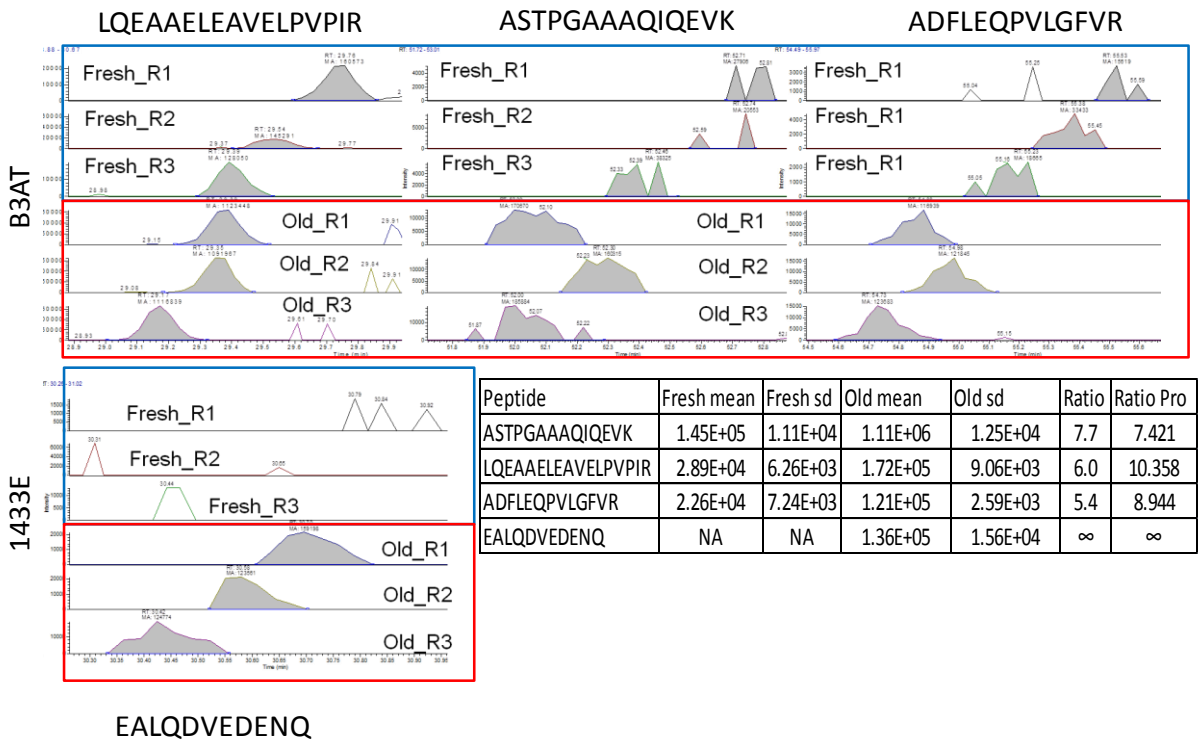


Figure 4: Principle component analysis of extracted features. PCA visualization is used to verify the efficiency of the filtered features (p<0.05 and fold-change > 5). Red and blue dots are the projection of data into scores, for old and fresh samples, respectively. The first component explains more than 89.20% of data variability whereas the second component explains 4.41%.



Peptide	Fresh mean	Fresh sd	Old mean	Old sd	Ratio	Ratio Pro
ASTPGAAAQIQEVK	1.45E+05	1.11E+04	1.11E+06	1.25E+04	7.7	7.421
LQEAAELEAVELPVPIR	2.89E+04	6.26E+03	1.72E+05	9.06E+03	6.0	10.358
ADFLEQPVLGFVR	2.26E+04	7.24E+03	1.21E+05	2.59E+03	5.4	8.944
EALQDVEDENQ	NA	NA	1.36E+05	1.56E+04	∞	∞

Figure 5: Extracted ion chromatogram of over-expressed peptides in old samples. Peptides from protein B3AT (D) and protein 1433E (E) are detected as over-expressed after comparison of extracted ion chromatograms. The ratio for each peptide between old and fresh samples is reported in the table. The values in the Ratio Pro column are from Progenesis

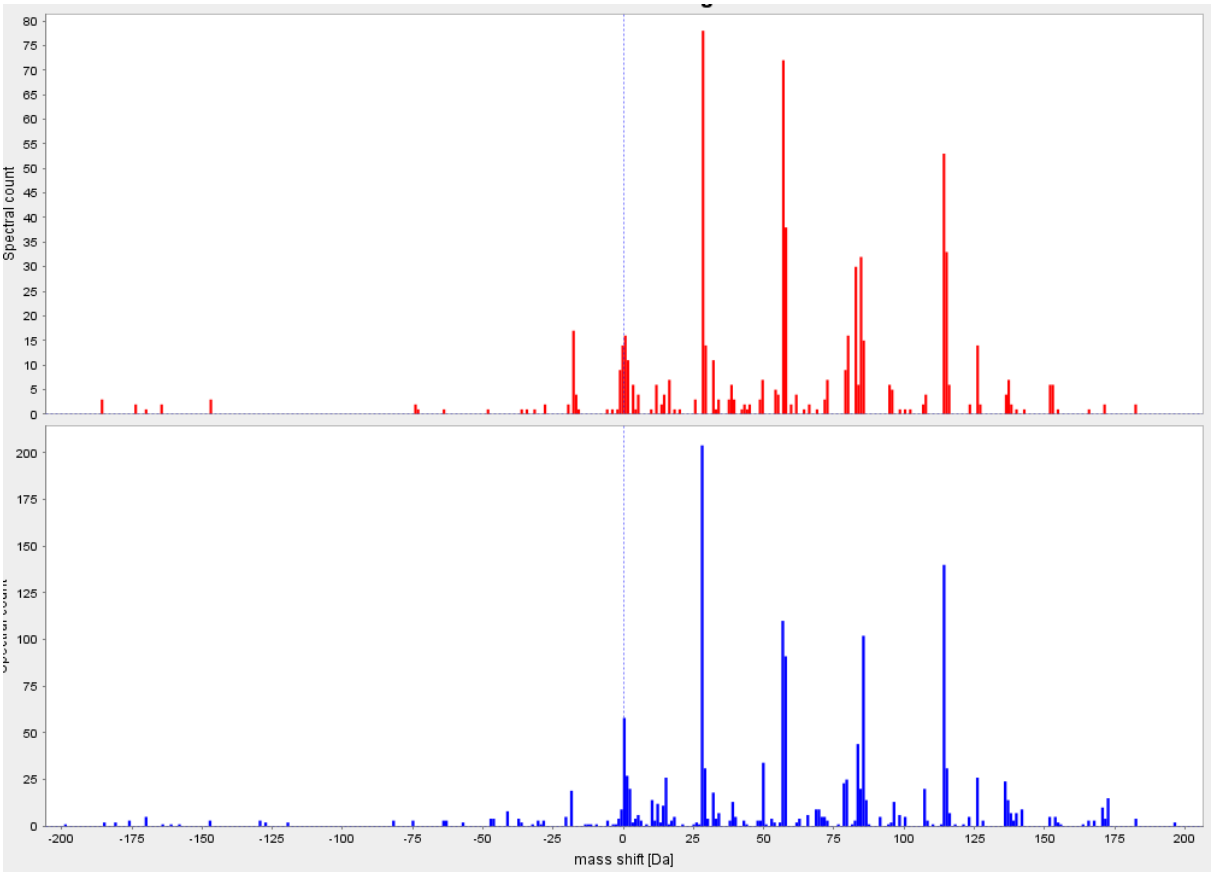


Figure 6: Integration of MS1 quantification with OMS data. The frequency of detected mass-shift from old (red) and fresh (blue) samples data is shown as histograms. OMS identified 666 and 1485 tandem mass spectra with mass-shifts between -200 and 200 Da from fresh and old samples, respectively.

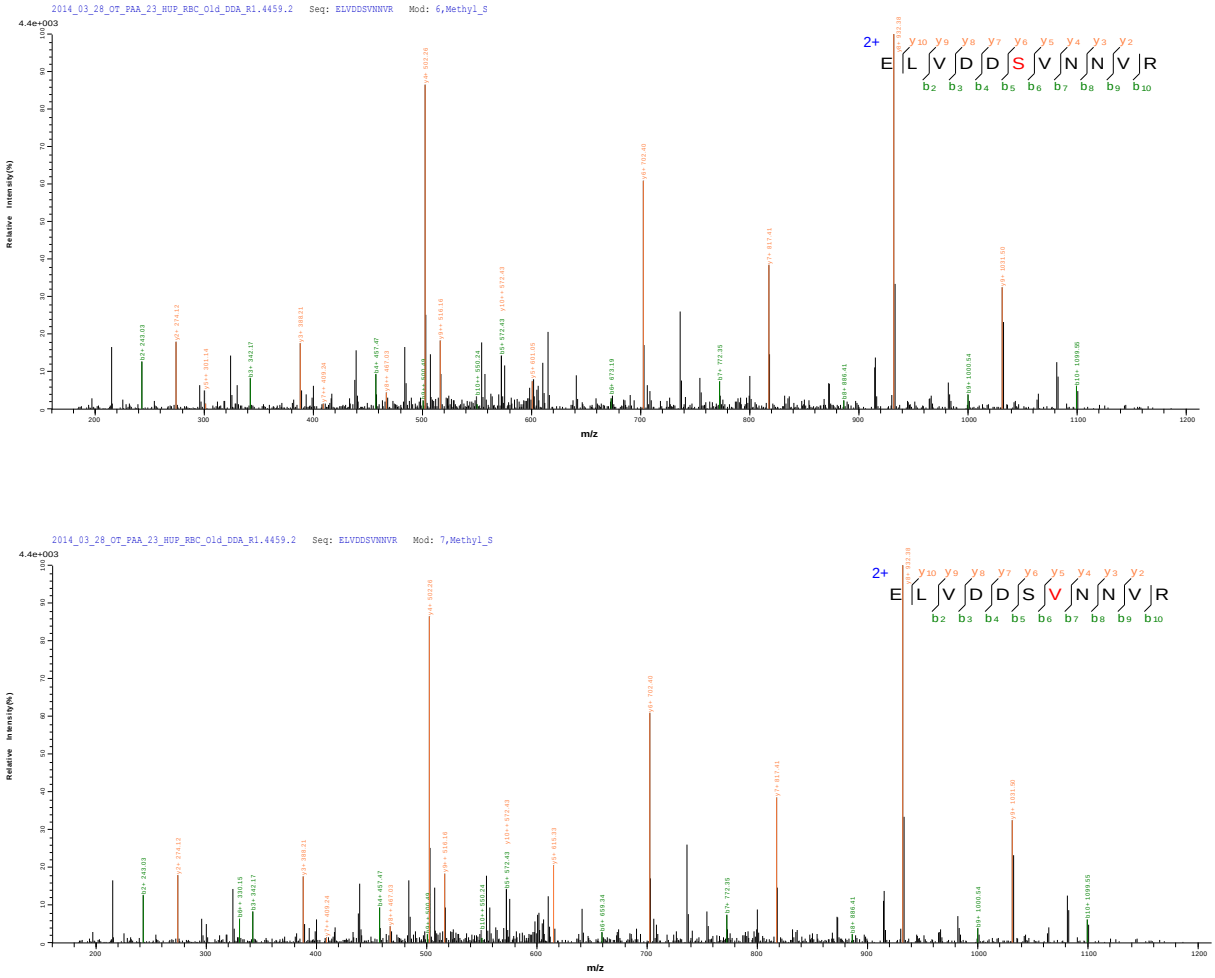


Figure 7: Tandem mass spectrum with methylation or substitution. The peptide ELVDDSVNNVR (PTIM1 protein) carries either a methylation on the S residue or a substitution of Valine into Isoleucine.

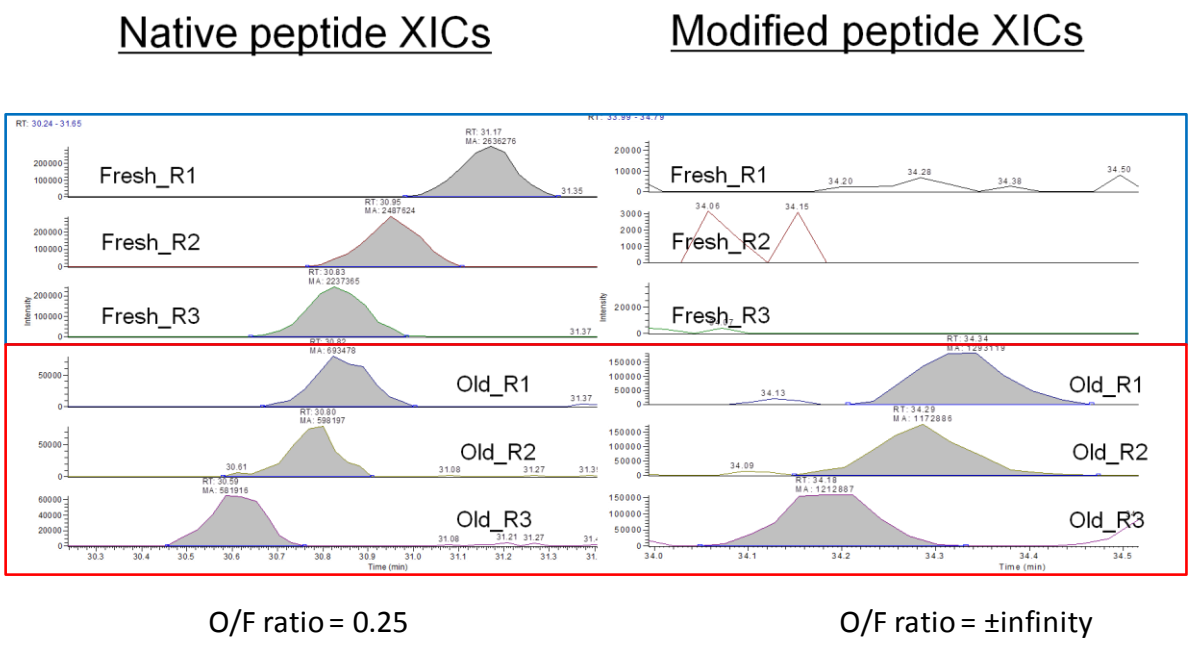


Figure 8: Extracted ion chromatogram of native and modified peptide ELVDDSVNNVR. Extracted ion chromatograms and AUC integration of the peptide ions are computed for each replicate and condition

5 Discussions

5.1 Common issues of data-independent acquisition

Data-independent tandem mass spectrometry was developed as an alternative method of precursor ion sampling compared to data-dependent tandem mass spectrometry. This development was necessary to overcome the biased selection towards abundant precursor ions and the redundant selection of the same precursor ion over the entire chromatographic elution. Dynamic exclusion algorithms, implemented in almost all mass spectrometers to prevent this redundant selection, greatly improved the dynamic range of detected peptides and proteins. However, this improvement still does not allow detection of very low abundance peptides in complex mixtures. In addition, data-dependent ion sampling methods incorporate a random component, limiting the reproducibility of precursor ion selection for multiple comparisons.

Purvine and co-workers opened the way of limiting the issues of DDA with an ion sampling method called "Shotgun-CID" implemented in an ESI-ToF mass spectrometer. It was the first attempt to select ions independent from their abundance. It consisted of parallel activation of all ions by in-source fragmentation and reconstitution of the parent-fragment ion lineage using the similarity of their chromatographic elution profile. Silva and co-workers improved this ion sampling method by acquiring alternatively low- and high-energy CID data in a Q-TOF hybrid instrument. Furthermore, they developed a suit of algorithms to process the data for identification and quantification. The advantages of Shotgun-CID-like ion sampling methods were the high-mass accuracy measurements, the preservation of ion chromatogram integrity and a duty cycle that could reach 100%. However, the complexity of the acquired multiplexed tandem mass spectra is increased due to the presence of product ions from multiple peptides. Even with high-mass accuracy measurements, their interpretations were difficult. This often led to the creation of sample-specific database to identify and quantify peptides and proteins.

Venable and co-workers proposed a solution to this issue, by reducing the width of the isolation window for parallel activation of precursor ions in nominal mass-resolution and

accuracy ion trap mass analyzer. They used an isolation window of 10 *m/z*, and all precursor ions within this window were activated prior to product ion scan. With such ion sampling methods, the resulting tandem mass spectra contained less chemical background and less contaminant product ions, thus increasing signal-to-noise ratio. In addition, they demonstrated that quantification based on tandem mass data was more accurate compared to simple scan data. Due to the large isolation window, the identification of tandem mass data was performed with a precursor ion tolerance above 10 Da. In order to preserve the ion chromatogram integrity, the idea of multiple injections was implicitly incorporated in the concept of Venable and co-worker.

Gillet and co-workers proposed an ion sampling method at the interface between shotgun-CID and the data-independent method proposed by Venable and co-workers, called SWATH. The method uses an isolation window of 25 Th in a Q-ToF mass analyzer. The advantage is high-mass resolution and accuracy measurement of ions for simple and tandem mass scans, and the acquisition of the entire data set in a single analysis. However, the method needs previously generated sample-specific libraries and a suit of algorithms for the identification and quantification of peptides. Consequently, a direct identification from SWATH data is not possible.

5.2 Data-independent acquisition using narrow isolation window.

Data-independent acquisition was often combined with high-mass accuracy and high-resolution analyzers and large precursor ion isolation windows to limit the number of analysis. However, data-independent acquisition was also implemented with mass analyzer operating at nominal mass accuracy and resolution, using relatively narrow precursor ion isolation windows. On the counterpart, it implied multiple injection of the same sample to cover the desired *m/z* mass range. Multiple analysis of the same sample is necessary to preserve chromatogram integrity for correct peptide/protein identifications and their potential quantification. Narrower precursor ion isolation windows limit the rate of peptide co-fragmentation and facilitate the interpretation and identification of tandem mass data. In average, an isolation window of four to five *m/z* units contains less than five precursor ions [144]. Based on this principle, the PAcIFIC method uses an isolation window comparable to the one used in data-dependent acquisition. Panchaud and co-workers measured the percentage of multiplexed tandem mass data to be 8%. This value is comparable to the rate of chimeric tandem mass data observed in data-dependent acquisition. One of the advantages of PAcIFIC over other data-independent acquisition methods is that the acquired tandem mass data can directly be submitted to database or library search for identification. Typically, a precursor ion mass tolerance of 3.75 Da was used for database search in the original description of PAcIFIC. Generally, precursor ion mass tolerance is the first filter used to reduce the number of tandem mass data to be matched with *in-silico* peptide fragmentation from databases. Thus, the precursor ion mass measurement accuracy and tolerance during database search is important to decrease data analysis time. However, with PAcIFIC and other DIA methods, this value is only approximatively known, *i.e.* it corresponds to the center of precursor ion isolation window. Two solutions have been proposed to measure or to calculate the precursor ion mass from multiplexed tandem mass data so far. The first one was proposed

by Carvahlo and co-worker, and consists in acquiring simple scans before serial acquisition of data-independent tandem mass data. The simple scans were used to detect potential precursor ions (see Chapter 1, section 1.6.1). However, it should be kept in mind that many high quality tandem mass spectra, matched to peptide sequences after database search, have no corresponding precursor ion signal detectable in the simple scans. These are the so-called orphan peptides. The proposed approach will thus not result with valid measurements for up to 30% of the acquired tandem mass spectra. The second solution was proposed by Venable and co-worker and used the complementarity of product ions from cleavage of the same peptide bond (see Chapter 1, section 1.6.3). The precursor ion *m/z* value can then be determined from the tandem mass spectrum with the same mass accuracy as the product ion measurement.

5.3 Application of PAcIFIC tandem mass spectrometry in proteomics.

In this thesis, we evaluated the use of narrow isolation window for data-independent acquisitions. First, we used a targeted approach called pseudo-multiple selected reaction monitoring (p-mSRM) to show the efficiency of relative quantification based on the systematic interrogation of predefined ion mass channels over the entire chromatographic elution time, regardless of ion abundance (see Chapter 2). The digested standard proteins were spiked into complex peptide mixtures (human plasma digests). Relative abundance was measured from AUC values from extracted product ion chromatograms for concentration ranging from 1 amol to 1 pmol on column, while keeping the plasma concentration constant We showed that it was possible to obtain linear responses over five orders of magnitude, from 10 amol to 1 pmol. This very high dynamic range is obtained from the combination of the dynamic range of ion detection in the linear ion trap and the dynamic range of the ion accumulation time into the trap. Indeed, the intra-spectrum dynamic range of an ion trap is

around 10^2-10^3 and the ion accumulation time varies between 0.1 to 100 ms. The total dynamic range can be estimated as being the product of these two values, *i.e.* 10^5. These results clearly showed the high ion sampling power of data-independent acquisition methods

In Chapter 3, we extended the application of p-mSRM ion sampling method to large-scale proteomics experiment, using PAcIFIC. Instead of targeting precursor ions, a list of fixed and predefined number of precursor ion channels were acquired independent from ion abundances. The number of acquired data-independent tandem mass spectra was limited to a cycle of 20 precursor ion channels per injection, in order to preserve the chromatogram integrity. A suit of algorithm was developed to process and to cluster tandem mass data, to separate co-eluting peptide and to re-calculate precursor ion *m/z* value based on product ion *m/z* value complementarity. As expected, PAcIFIC ion sampling yielded larger number of peptide/protein identifications compared to data-dependent ion sampling methods. Clustering of tandem mass spectra improved the peptide identification rate by 8 %, and reduced the spectra to be searched by 18 %. In addition, the size of the clustered data decreased by 55 %. In addition, the number of unique peptide identifications increased by 14 % after re-calculation of precursor ions. The counts of chimeric spectra containing two or more spectra showed that multiple hits per spectrum were reduced after data processing. This confirmed that the clustering algorithm was able to separate chimeric spectra based on the chromatographic elution profile of product ions. The method used for precursor ion mass re-calculation is very similar to the one described by Venable and co-workers [152]. As reported by them, this method based on fragment ions complementarity is affected by the use of large isolation window. The narrow precursor ion isolation window used for PAcIFIC is thus particularly suited for this approach.

In chapter 4, PAcIFIC and data-dependent acquisitions were combined for a comprehensive analysis of the proteome of aging red blood cells. In this study, the goal was to identify

possible peptide modifications, including chemical adducts, PTMs and other possible modifications related to storage bag conservation and RBC aging. For this, we used an open modification search approach using previously generated spectral libraries. In summary, we identified more peptides and proteins from PAcIFIC data compared to DDA. We observed less peptide identifications in old samples compared to fresh samples, indicating an effect of possible storage lesion of RBCs during the prolonged duration in blood storage bags. In order to extract and to reduce the number of possible peptide candidates that could indicate RBC aging, we performed label-free quantification from data-dependent acquisitions. As reported in chapter 4, more than 13000 features were extracted and filtered using one-way ANOVA analysis. This value was then reduced to 3028 features if only differentially abundant signals were considered. Finally, 13 proteins corresponding to 17 unique proteotypique peptides were matched with these features. The manual extraction of ion chromatograms for AUC measurements of these 17 peptides showed good agreement with the automatically generated data using label-free quantification software. Open modification search was then performed using a spectral library generated with the previously identified peptides. Using the same filtering approaches, the number of tandem mass data considered as modified peptides was reduced from 14000 to 1797. Among them, 555 tandem mass spectra were detected in fresh samples and 1242 tandem mass spectra were detected in old samples. The data was then filtered to keep only proteins previously reported to be part of RBCs. All possible non-relevant modifications potentially occurring during sample preparation, such as carbamidomethylation and di-carbamidomethylation and others were discarded. Finally, five modified peptides, corresponding to protein Carbonic anhydrase 1, Catalase, and PTIM were identified in old samples. All reported modification mass-shifts were within 14 to 30 Da. In particular, peptide ELVDDSVNNVR of protein PTIM that carries a mass-shift of 14.04 Da either on the S residue, corresponding to a methylation, or on the V residue, corresponding to

an amino acid substitution into an I residue. The amino acid substitution of V into I was previously described in genetic studies and corresponded to an allele frequency of 0.45 in Caucasian populations [189]. We confirmed that the mass-shift probably corresponded to the substitution of V into I by performing manual annotation of the tandem mass spectrum. The native peptide of protein PTIM has higher abundance in fresh samples compared to old samples. In contrast, the variant peptide is more abundant in old RBC samples. This difference could be related to stability and degradation susceptibility as reported by DeIvry and co-workers [189]. The identified modified peptides could potentially serve as marker for RBC aging in blood storage bags.

6 Conclusion and perspectives

6.1 Conclusions

Data-independent acquisition is now widely used in proteomics and metabolomics. The reason of this success is the advantage of systematic interrogation and isolation of all precursor ions, regardless of their abundance. This way of performing ion sampling improves detection of low abundance species, data reproducibility, and allows acquiring a large number of tandem mass spectra. The drawbacks of the data-independent acquisitions are that the precursor ion isolation window is not centered on the peptide *m/z* for tandem mass spectrometry, and that the rate of chimeric tandem mass spectra is increased. This induces inexactitude during database search and can lead to false-positive identifications. An adapted strategy for the processing of tandem mass spectra acquired in data-independent mode and for the separation of chimeric spectra is presented here.

This suit of algorithms was applied to PAcIFIC tandem mass spectrometry data. The application of these algorithms improved the number of identified peptides. In addition, the number of chimeric spectra was reduced, showing the decomposition of multiplexed tandem mass spectra into spectrum containing only product ions from one precursor ion. The method was then applied for a comprehensive analysis of the red blood cell proteome, and in particular, to identify modified peptides during RBC aging. Modified peptides appearing only in old RBC samples were identified. They are potential markers of RBC aging in blood storage bags.

6.2 Perspectives

Data-independent acquisition is the alternative ion sampling method to improve the selectivity and sensitivity of low abundance ion selection, number of peptide identifications and the robustness of data reproducibility. However, the use of large isolation windows, not centered on the peptide ion *m/z* value, requires extensive data processing and creation of sample spectral libraries for peptide identification and quantification. In contrast to narrow isolation windows, the scan frequency of tandem mass spectra acquisition is limited to ten per seconds [144], leading to an increase of total analysis time, by performing multiple injections of the same sample.

The main issue with DIA acquisitions with narrow isolation windows is instrument acquisition speed, to limit the number of repetitive injections. In this regard, Houle Wang and co-workers proposed a hybrid mass spectrometer able to perform tandem mass spectrometry on all precursor ions sequentially exiting an ion trap [190]. The geometry of the system allows to conduct two stages of mass analysis: The ion trap is used for pulsed, mass selective ejection of precursor ions prior to activation and time of flight analysis. A duty cycle of 100% is obtained because all accumulated precursor ions leave the trap as ion packets of small *m/z* range prior to tandem mass spectrometry. Such an instrument design demonstrates that high frequency acquisition of tandem mass spectra for DIA is possible. However, the amount of data generated is enormous. Algorithms to reduce the data and to increase their accuracy are thus essential. Such algorithms are presented here.

7 References

[1] R. Aebersold and D. R. Goodlett, "Mass Spectrometry in Proteomics," *Chem. Rev.*, vol. 101, no. 2, pp. 269–296, Feb. 2001.
[2] P. H. O'Farrell, "High Resolution Two-Dimensional Electrophoresis of Proteins," *J. Biol. Chem.*, vol. 250, no. 10, pp. 4007–4021, May 1975.
[3] P. Edman, E. Högfeldt, L. G. Sillén, and P.-O. Kinell, "Method for Determination of the Amino Acid Sequence in Peptides.," *Acta Chem. Scand.*, vol. 4, pp. 283–293, 1950.
[4] H. Towbin, T. Staehelin, and J. Gordon, "Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 76, no. 9, pp. 4350–4354, Sep. 1979.
[5] J. Renart, J. Reiser, and G. R. Stark, "Transfer of proteins from gels to diazobenzyloxymethyl-paper and detection with antisera: a method for studying antibody specificity and antigen structure.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 76, no. 7, pp. 3116–3120, Jul. 1979.
[6] F. Gerber, M. Krummen, H. Potgeter, A. Roth, C. Siffrin, and C. Spoendlin, "Practical aspects of fast reversed-phase high-performance liquid chromatography using 3 μm particle packed columns and monolithic columns in pharmaceutical development and production working under current good manufacturing practice," *J. Chromatogr. A*, vol. 1036, no. 2, pp. 127–133, May 2004.
[7] P. Arpino, M. A. Baldwin, and F. W. McLafferty, "Liquid chromatography-mass spectrometry. II—continuous monitoring," *Biol. Mass Spectrom.*, vol. 1, no. 1, pp. 80–82, Feb. 1974.
[8] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, "Electrospray ionization for mass spectrometry of large biomolecules," *Science*, vol. 246, no. 4926, pp. 64–71, Oct. 1989.
[9] N. M. Griffin, J. Yu, F. Long, P. Oh, S. Shore, Y. Li, J. A. Koziol, and J. E. Schnitzer, "Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis," *Nat. Biotechnol.*, vol. 28, no. 1, pp. 83–89, Jan. 2010.
[10] Z. Li, R. M. Adams, K. Chourey, G. B. Hurst, R. L. Hettich, and C. Pan, "Systematic Comparison of Label-Free, Metabolic Labeling, and Isobaric Chemical Labeling for Quantitative Proteomics on LTQ Orbitrap Velos," *J. Proteome Res.*, vol. 11, no. 3, pp. 1582–1590, Mar. 2012.
[11] J. F. de la Mora, G. J. Van Berkel, C. G. Enke, R. B. Cole, M. Martinez-Sanchez, and J. B. Fenn, "Electrochemical processes in electrospray ionization mass spectrometry," *J. Mass Spectrom.*, vol. 35, no. 8, pp. 939–952, Aug. 2000.
[12] T. C. Rohner, N. Lion, and H. H. Girault, "Electrochemical and theoretical aspects of electrospray ionisation," *Phys. Chem. Chem. Phys.*, vol. 6, no. 12, pp. 3056–3068, Jun. 2004.
[13] S. Purvine, J.-T. Eppel*, E. C. Yi, and D. R. Goodlett, "Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer," *PROTEOMICS*, vol. 3, no. 6, pp. 847–850, 2003.
[14] J. C. Silva, R. Denny, C. A. Dorschel, M. Gorenstein, I. J. Kass, G.-Z. Li, T. McKenna, M. J. Nold, K. Richardson, P. Young, and S. Geromanos, "Quantitative Proteomic Analysis by Accurate Mass Retention Time Pairs," *Anal. Chem.*, vol. 77, no. 7, pp. 2187–2200, Apr. 2005.
[15] A. Panchaud, A. Scherl, S. A. Shaffer, P. D. von Haller, H. D. Kulasekara, S. I. Miller, and D. R. Goodlett, "Precursor Acquisition Independent From Ion Count: How to Dive Deeper into the Proteomics Ocean," *Anal. Chem.*, vol. 81, no. 15, pp. 6481–6488, Aug. 2009.
[16] V. Lange, P. Picotti, B. Domon, and R. Aebersold, "Selected reaction monitoring for quantitative proteomics: a tutorial," *Mol. Syst. Biol.*, vol. 4, no. 1, p. 222, Jan. 2008.
[17] J.-H. Baek, H. Kim, B. Shin, and M.-H. Yu, "Multiple Products Monitoring as a Robust Approach for Peptide Quantification," *J. Proteome Res.*, vol. 8, no. 7, pp. 3625–3632, Jul. 2009.
[18] H. Pak, C. Pasquarello, and A. Scherl, "Label-free protein quantification on tandem mass spectra acquired in a data-independent mode provides accurate measurements over five orders of concentration magnitude in complex matrices," *J. Integr. OMICS*, vol. 1, no. 2, Dec. 2011.

[19] V. R. Meyer, *Practical High-Performance Liquid Chromatography*. John Wiley & Sons, 2010.
[20] H. D. Meiring, E. van der Heeft, G. J. ten Hove, and A. P. J. M. de Jong, "Nanoscale LC–MS(n): technical design and applications to peptide and protein analysis," *J. Sep. Sci.*, vol. 25, no. 9, pp. 557–568, Jun. 2002.
[21] A. J. Link, J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik, and J. R. Yates, "Direct analysis of protein complexes using mass spectrometry," *Nat. Biotechnol.*, vol. 17, no. 7, pp. 676–682, juillet 1999.
[22] A. Ducret, I. V. Oostveen, J. K. Eng, J. R. Yates, and R. Aebersold, "High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry," *Protein Sci.*, vol. 7, no. 3, pp. 706–719, Mar. 1998.
[23] D. A. Wolters, M. P. Washburn, and J. R. Yates, "An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics," *Anal. Chem.*, vol. 73, no. 23, pp. 5683–5690, décembre 2001.
[24] C. L. Gatlin, J. K. Eng, S. T. Cross, J. C. Detter, and J. R. Yates 3rd, "Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry," *Anal. Chem.*, vol. 72, no. 4, pp. 757–763, Feb. 2000.
[25] M. R. Emmett and R. M. Caprioli, "Micro-electrospray mass spectrometry: ultra-high-sensitivity analysis of peptides and proteins," *J. Am. Soc. Mass Spectrom.*, vol. 5, no. 7, pp. 605–613, Jul. 1994.
[26] L. J. Licklider, C. C. Thoreen, J. Peng, and S. P. Gygi, "Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column," *Anal. Chem.*, vol. 74, no. 13, pp. 3076–3083, Jul. 2002.
[27] E. C. Yi, H. Lee, R. Aebersold, and D. R. Goodlett, "A microcapillary trap cartridge-microcapillary high-performance liquid chromatography electrospray ionization emitter device capable of peptide tandem mass spectrometry at the attomole level on an ion trap mass spectrometer with automated routine operation," *Rapid Commun. Mass Spectrom. RCM*, vol. 17, no. 18, pp. 2093–2098, 2003.
[28] K. Wanner and G. H?fner, *Mass Spectrometry in Medicinal Chemistry: Applications in Drug Discovery*. John Wiley & Sons, 2007.
[29] E. de Hoffmann and V. Stroobant, *Mass Spectrometry: Principles and Applications*. John Wiley & Sons, 2007.
[30] A. G. Marshall, C. L. Hendrickson, and S. D. H. Shi, "Scaling MS plateaus with high-resolution FT-ICRMS," *Anal. Chem.*, vol. 74, no. 9, p. 252A–259A, May 2002.
[31] R. Zubarev and M. Mann, "On the proper use of mass accuracy in proteomics," *Mol. Cell. Proteomics MCP*, vol. 6, no. 3, pp. 377–381, Mar. 2007.
[32] W. Bleakney, "The Ionization of Hydrogen by Single Electron Impact," *Phys. Rev.*, vol. 35, no. 10, pp. 1180–1186, May 1930.
[33] A. G. Harrison, *Chemical Ionization Mass Spectrometry, Second Edition*. CRC Press, 1992.
[34] M. Karas, D. Bachmann, U. Bahr, and F. Hillenkamp, "Matrix-assisted ultraviolet laser desorption of non-volatile compounds," *Int. J. Mass Spectrom. Ion Process.*, vol. 78, pp. 53–68, Sep. 1987.
[35] K. Strupat, M. Karas, and F. Hillenkamp, "2,5-Dihydroxybenzoic acid: a new matrix for laser desorption--ionization mass spectrometry," *Int. J. Mass Spectrom. Ion Process.*, vol. 111, pp. 89–102, Dec. 1991.
[36] E. C. Huang and J. D. Henion, "Packed-capillary liquid chromatography/ion-spray tandem mass spectrometry determination of biomolecules," *Anal. Chem.*, vol. 63, no. 7, pp. 732–739, Apr. 1991.
[37] A. P. Bruins, T. R. Covey, and J. D. Henion, "Ion spray interface for combined liquid chromatography/atmospheric pressure ionization mass spectrometry," *Anal. Chem.*, vol. 59, no. 22, pp. 2642–2646, Nov. 1987.
[38] P. Kebarle, "A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry," *J. Mass Spectrom. JMS*, vol. 35, no. 7, pp. 804–817, Jul. 2000.

[39] Covey Thomas, "Analytical Characteristics of the Electrospray Ionization Process," in *Biochemical and Biotechnological Applications of Electrospray Ionization Mass Spectrometry*, vol. 619, 0 vols., American Chemical Society, 1996, pp. 21–59.
[40] W. Paul and H. Steinwedel, "Ein neues Massenspektrometer ohne Magnetfeld," *Z. Naturforschung Teil A*, vol. 8, p. 448, 1953.
[41] P. E. Miller and M. B. Denton, "The quadrupole mass filter: Basic operating concepts," *J. Chem. Educ.*, vol. 63, no. 7, p. 617, Jul. 1986.
[42] R. E. March, A. W. McMahon, F. A. Londry, R. L. Alfred, J. F. J. Todd, and F. Vedel, "Resonance excitation of ions stored in a quadrupole ion trap. Part 1. A simulation study," *Int. J. Mass Spectrom. Ion Process.*, vol. 95, no. 2, pp. 119–156, Dec. 1989.
[43] W. Paul, "Electromagnetic Traps for Charged and Neutral Particles (Nobel Lecture)," *Angew. Chem. Int. Ed. Engl.*, vol. 29, no. 7, pp. 739–748, Jul. 1990.
[44] R. E. March, "An Introduction to Quadrupole Ion Trap Mass Spectrometry," *J. Mass Spectrom.*, vol. 32, no. 4, pp. 351–369, Apr. 1997.
[45] R. E. March and J. F. J. Todd, *Practical Aspects of Trapped Ion Mass Spectrometry: Applications of Ion Trapping Devices*. CRC Press, 2009.
[46] D. J. Douglas, A. J. Frank, and D. Mao, "Linear ion traps in mass spectrometry," *Mass Spectrom. Rev.*, vol. 24, no. 1, pp. 1–29, Feb. 2005.
[47] J. W. Hager, "A new linear ion trap mass spectrometer," *Rapid Commun. Mass Spectrom.*, vol. 16, no. 6, pp. 512–526, Mar. 2002.
[48] D. C. Imrie, J. M. Pentney, and J. S. Cottrell, "A Faraday cup detector for high-mass ions in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry," *Rapid Commun. Mass Spectrom.*, vol. 9, no. 13, pp. 1293–1296, Jan. 1995.
[49] P. Önnerfjord, J. Nilsson, L. Wallman, T. Laurell, and G. Marko-Varga, "Picoliter Sample Preparation in MALDI-TOF MS Using a Micromachined Silicon Flow-Through Dispenser," *Anal. Chem.*, vol. 70, no. 22, pp. 4755–4760, Nov. 1998.
[50] B. A. Mamyrin, V. I. Karataev, D. V. Shmikk, and V. A. Zagulin, "The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution," *Sov. J. Exp. Theor. Phys.*, vol. 37, p. 45, Jul. 1973.
[51] J. H. J. Dawson and M. Guilhaus, "Orthogonal-acceleration time-of-flight mass spectrometer," *Rapid Commun. Mass Spectrom.*, vol. 3, no. 5, pp. 155–159, May 1989.
[52] J. A. Hipple, H. Sommer, and H. A. Thomas, "A Precise Method of Determining the Faraday by Magnetic Resonance," *Phys. Rev.*, vol. 76, no. 12, pp. 1877–1878, Dec. 1949.
[53] M. B. Comisarow and A. G. Marshall, "Fourier transform ion cyclotron resonance spectroscopy," *Chem. Phys. Lett.*, vol. 25, no. 2, pp. 282–283, Mar. 1974.
[54] "Makarov, A. (1999) Mass spectrometer. US Patent, 5886346."
[55] A. Makarov, "Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis," *Anal. Chem.*, vol. 72, no. 6, pp. 1156–1162, Mar. 2000.
[56] A. Makarov, E. Denisov, O. Lange, and S. Horning, "Dynamic Range of Mass Accuracy in LTQ Orbitrap Hybrid Mass Spectrometer," *J. Am. Soc. Mass Spectrom.*, vol. 17, no. 7, pp. 977–982, Jul. 2006.
[57] G. L. Glish, S. A. McLuckey, E. H. McBay, and L. K. Bertram, "Design and performance of a hybrid mass spectrometer of QEB geometry," *Int. J. Mass Spectrom. Ion Process.*, vol. 70, no. 3, pp. 321–338, Jul. 1986.
[58] V. M. Doroshenko and R. J. Cotter, "A quadrupole ion trap/time-of-flight mass spectrometer with a parabolic reflectron," *J. Mass Spectrom. JMS*, vol. 33, no. 4, pp. 305–318, Apr. 1998.
[59] H. R. Morris, T. Paxton, A. Dell, J. Langhorne, M. Berg, R. S. Bordoli, J. Hoyes, and R. H. Bateman, "High Sensitivity Collisionally-activated Decomposition Tandem Mass Spectrometry on a Novel Quadrupole/Orthogonal-acceleration Time-of-flight Mass Spectrometer," *Rapid Commun. Mass Spectrom.*, vol. 10, no. 8, pp. 889–896, Jan. 1996.
[60] J. C. Schwartz, A. P. Wade, C. G. Enke, and R. G. Cooks, "Systematic delineation of scan modes in multidimensional mass spectrometry," *Anal. Chem.*, vol. 62, no. 17, pp. 1809–1818, Sep. 1990.

[61] A. G. Harrison and M. S. Lin, "Energy dependence of the fragmentation of the n-butylbenzene molecular ion," *Int. J. Mass Spectrom. Ion Phys.*, vol. 51, no. 2–3, pp. 353–356, Jul. 1983.
[62] H. Yamaoka, P. Dông, and J. Durup, "Energetics of the Collision-Induced Dissociations C2H2+→C2H+ + H and C2H2+→H+ + C2H," *J. Chem. Phys.*, vol. 51, no. 8, pp. 3465–3476, Oct. 1969.
[63] R. N. Schwartz, Z. I. Slawsky, and K. F. Herzfeld, "Calculation of Vibrational Relaxation Times in Gases," *J. Chem. Phys.*, vol. 20, no. 10, pp. 1591–1599, Oct. 1952.
[64] M. A. Mabud, M. J. Dekrey, and R. Graham Cooks, "Surface-induced dissociation of molecular ions," *Int. J. Mass Spectrom. Ion Process.*, vol. 67, no. 3, pp. 285–294, Nov. 1985.
[65] V. Grill, J. Shen, C. Evans, and R. G. Cooks, "Collisions of ions with surfaces at chemically relevant energies: Instrumentation and phenomena," *Rev. Sci. Instrum.*, vol. 72, pp. 3149–3179, Aug. 2001.
[66] D. P. Little, J. P. Speir, M. W. Senko, P. B. O'Connor, and F. W. McLafferty, "Infrared Multiphoton Dissociation of Large Multiply Charged Ions for Biomolecule Sequencing," *Anal. Chem.*, vol. 66, no. 18, pp. 2809–2815, Sep. 1994.
[67] R. A. Zubarev, N. L. Kelleher, and F. W. McLafferty, "Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process," *J. Am. Chem. Soc.*, vol. 120, no. 13, pp. 3265–3266, Apr. 1998.
[68] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt, "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 26, pp. 9528–9533, Jun. 2004.
[69] P. Roepstorff and J. Fohlman, "Letter to the editors," *Biol. Mass Spectrom.*, vol. 11, no. 11, pp. 601–601, Nov. 1984.
[70] R. S. Johnson, S. A. Martin, K. Biemann, J. T. Stults, and J. T. Watson, "Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine," *Anal. Chem.*, vol. 59, no. 21, pp. 2621–2625, Nov. 1987.
[71] K. Biemann, "Contributions of mass spectrometry to peptide and protein structure," *Biomed. Environ. Mass Spectrom.*, vol. 16, no. 1–12, pp. 99–111, Oct. 1988.
[72] W. D. van Dongen, H. F. Ruijters, H. J. Luinge, W. Heerma, and J. Haverkamp, "Statistical analysis of mass spectral data obtained from singly protonated peptides under high-energy collision-induced dissociation conditions," *J. Mass Spectrom. JMS*, vol. 31, no. 10, pp. 1156–1162, Oct. 1996.
[73] N. L. Anderson and N. G. Anderson, "Proteome and proteomics: New technologies, new concepts, and new words," *ELECTROPHORESIS*, vol. 19, no. 11, pp. 1853–1861, Aug. 1998.
[74] M. R. Wilkins, C. Pasquali, R. D. Appel, K. Ou, O. Golaz, J.-C. Sanchez, J. X. Yan, A. A. Gooley, G. Hughes, I. Humphery-Smith, K. L. Williams, and D. F. Hochstrasser, "From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Arnino Acid Analysis," *Nat. Biotechnol.*, vol. 14, no. 1, pp. 61–65, Jan. 1996.
[75] A. Moradian, A. Kalli, M. J. Sweredoski, and S. Hess, "The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications," *PROTEOMICS*, vol. 14, no. 4–5, pp. 489–497, Mar. 2014.
[76] L. M. Smith, N. L. Kelleher, and T. C. for T. D. Proteomics, "Proteoform: a single term describing protein complexity," *Nat. Methods*, vol. 10, no. 3, pp. 186–187, Mar. 2013.
[77] J. Colinge, A. Masselot, M. Giron, T. Dessingy, and J. Magnin, "OLAV: towards high-throughput tandem mass spectrometry data identification," *Proteomics*, vol. 3, no. 8, pp. 1454–1463, Aug. 2003.
[78] J. R. Yates, S. F. Morgan, C. L. Gatlin, P. R. Griffin, and J. K. Eng, "Method To Compare Collision-Induced Dissociation Spectra of Peptides: Potential for Library Searching and Subtractive Analysis," *Anal. Chem.*, vol. 70, no. 17, pp. 3557–3565, Sep. 1998.

[79] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein, and R. Aebersold, "Development and validation of a spectral library searching method for peptide identification from MS/MS," *PROTEOMICS*, vol. 7, no. 5, pp. 655–667, 2007.

[80] R. Craig, J. C. Cortens, D. Fenyo, and R. C. Beavis, "Using Annotated Peptide Mass Spectrum Libraries for Protein Identification," *J. Proteome Res.*, vol. 5, no. 8, pp. 1843–1849, Aug. 2006.

[81] M. C. Wiener, J. R. Sachs, E. G. Deyanova, and N. A. Yates, "Differential Mass Spectrometry: A Label-Free LC–MS Method for Finding Significant Differences in Complex Peptide and Protein Mixtures," *Anal. Chem.*, vol. 76, no. 20, pp. 6085–6096, Oct. 2004.

[82] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, "Quantitative mass spectrometry in proteomics: a critical review," *Anal. Bioanal. Chem.*, vol. 389, no. 4, pp. 1017–1031, Oct. 2007.

[83] A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, and C. Hamon, "Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS," *Anal. Chem.*, vol. 75, no. 8, pp. 1895–1904, Apr. 2003.

[84] G. C. McAlister, E. L. Huttlin, W. Haas, L. Ting, M. P. Jedrychowski, J. C. Rogers, K. Kuhn, I. Pike, R. A. Grothe, J. D. Blethrow, and S. P. Gygi, "Increasing the Multiplexing Capacity of TMTs Using Reporter Ion Isotopologues with Isobaric Masses," *Anal. Chem.*, vol. 84, no. 17, pp. 7469–7478, Sep. 2012.

[85] J. Listgarten and A. Emili, "Statistical and Computational Methods for Comparative Proteomic Profiling Using Liquid Chromatography-Tandem Mass Spectrometry," *Mol. Cell. Proteomics*, vol. 4, no. 4, pp. 419–434, Apr. 2005.

[86] J. Gao, G. J. Opiteck, M. S. Friedrichs, A. R. Dongre, and S. A. Hefta, "Changes in the Protein Expression of Yeast as a Function of Carbon Source," *J. Proteome Res.*, vol. 2, no. 6, pp. 643–649, Dec. 2003.

[87] B. Zybailov, M. K. Coleman, L. Florens, and M. P. Washburn, "Correlation of Relative Abundance Ratios Derived from Peptide Ion Chromatograms and Spectrum Counting for Quantitative Proteomic Analysis Using Stable Isotope Labeling," *Anal. Chem.*, vol. 77, no. 19, pp. 6218–6224, Oct. 2005.

[88] B. Zybailov, A. L. Mosley, M. E. Sardiu, M. K. Coleman, L. Florens, and M. P. Washburn, "Statistical Analysis of Membrane Proteome Expression Changes in Saccharomyces cerevisiae," *J. Proteome Res.*, vol. 5, no. 9, pp. 2339–2347, Sep. 2006.

[89] L. Florens, M. J. Carozza, S. K. Swanson, M. Fournier, M. K. Coleman, J. L. Workman, and M. P. Washburn, "Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors," *Methods*, vol. 40, no. 4, pp. 303–311, Dec. 2006.

[90] H. Liu, R. G. Sadygov, and J. R. Yates, "A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics," *Anal. Chem.*, vol. 76, no. 14, pp. 4193–4201, juillet 2004.

[91] W. Zhou, L. A. Liotta, and E. F. Petricoin, "The Spectra Count Label-free Quantitation in Cancer Proteomics," *Cancer Genomics - Proteomics*, vol. 9, no. 3, pp. 135–142, May 2012.

[92] Y. Zhang, Z. Wen, M. P. Washburn, and L. Florens, "Effect of Dynamic Exclusion Duration on Spectral Count Based Quantitative Proteomics," *Anal. Chem.*, vol. 81, no. 15, pp. 6317–6326, Aug. 2009.

[93] M. J. MacCoss, C. C. Wu, H. Liu, R. Sadygov, and J. R. Yates, "A Correlation Algorithm for the Automated Quantitative Analysis of Shotgun Proteomics Data," *Anal. Chem.*, vol. 75, no. 24, pp. 6912–6921, Dec. 2003.

[94] X. Li, E. C. Yi, C. J. Kemp, H. Zhang, and R. Aebersold, "A Software Suite for the Generation and Comparison of Peptide Arrays from Sets of Data Collected by Liquid Chromatography-Mass Spectrometry," *Mol. Cell. Proteomics*, vol. 4, no. 9, pp. 1328–1340, Sep. 2005.

[95] W. Wang, H. Zhou, H. Lin, S. Roy, T. A. Shaler, L. R. Hill, S. Norton, P. Kumar, M. Anderle, and C. H. Becker, "Quantification of Proteins and Metabolites by Mass Spectrometry without Isotopic Labeling or Spiked Standards," *Anal. Chem.*, vol. 75, no. 18, pp. 4818–4826, Sep. 2003.

[96] R. E. Higgs, M. D. Knierman, V. Gelfanova, J. P. Butler, and J. E. Hale, "Comprehensive Label-Free Method for the Relative Quantification of Proteins from Biological Samples," *J. Proteome Res.*, vol. 4, no. 4, pp. 1442–1450, Aug. 2005.

[97] L. N. Mueller, M.-Y. Brusniak, D. R. Mani, and R. Aebersold, "An Assessment of Software Solutions for the Analysis of Mass Spectrometry Based Quantitative Proteomics Data," *J. Proteome Res.*, vol. 7, no. 1, pp. 51–61, Jan. 2008.

[98] L. N. Mueller, O. Rinner, A. Schmidt, S. Letarte, B. Bodenmiller, M.-Y. Brusniak, O. Vitek, R. Aebersold, and M. Müller, "SuperHirn – a novel tool for high resolution LC-MS-based peptide/protein profiling," *PROTEOMICS*, vol. 7, no. 19, pp. 3470–3480, Oct. 2007.

[99] M. Dakna, Z. He, W. C. Yu, H. Mischak, and W. Kolch, "Technical, bioinformatical and statistical aspects of liquid chromatography–mass spectrometry (LC–MS) and capillary electrophoresis-mass spectrometry (CE-MS) based clinical proteomics: A critical assessment," *J. Chromatogr. B*, vol. 877, no. 13, pp. 1250–1258, May 2009.

[100] V. P. Andreev, L. Li, L. Cao, Y. Gu, T. Rejtar, S.-L. Wu, and B. L. Karger, "A New Algorithm Using Cross-Assignment for Label-Free Quantitation with LC–LTQ-FT MS," *J. Proteome Res.*, vol. 6, no. 6, pp. 2186–2194, juin 2007.

[101] P. L. Courchesne, M. D. Jones, J. H. Robinson, C. S. Spahr, S. McCracken, D. L. Bentley, R. Luethy, and S. D. Patterson, "Optimization of capillary chromatography ion trap-mass spectrometry for identification of gel-separated proteins," *ELECTROPHORESIS*, vol. 19, no. 6, pp. 956–967, 1998.

[102] M. P. Washburn, D. Wolters, and J. R. Yates, "Large-scale analysis of the yeast proteome by multidimensional protein identification technology," *Nat. Biotechnol.*, vol. 19, no. 3, pp. 242–247, Mar. 2001.

[103] S. Hanash, "HUPO initiatives relevant to clinical proteomics," *Mol. Cell. Proteomics MCP*, vol. 3, no. 4, pp. 298–301, Apr. 2004.

[104] Y. Shen, J. M. Jacobs, D. G. Camp, R. Fang, R. J. Moore, R. D. Smith, W. Xiao, R. W. Davis, and R. G. Tompkins, "Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome," *Anal. Chem.*, vol. 76, no. 4, pp. 1134–1144, Feb. 2004.

[105] P. E. Michel, F. Reymond, I. L. Arnaud, J. Josserand, H. H. Girault, and J. S. Rossier, "Protein fractionation in a multicompartment device using Off-Gel™ isoelectric focusing," *ELECTROPHORESIS*, vol. 24, no. 1–2, pp. 3–11, Jan. 2003.

[106] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, Mar. 2003.

[107] T. Fortin, A. Salvador, J. P. Charrier, C. Lenz, F. Bettsworth, X. Lacoux, G. Choquet-Kastylevsky, and J. Lemoine, "Multiple Reaction Monitoring Cubed for Protein Quantification at the Low Nanogram/Milliliter Level in Nondepleted Human Serum," *Anal. Chem.*, vol. 81, no. 22, pp. 9343–9352, Nov. 2009.

[108] S. Yang, J. Cha, and K. Carlson, "Quantitative determination of trace concentrations of tetracycline and sulfonamide antibiotics in surface water using solid-phase extraction and liquid chromatography/ion trap tandem mass spectrometry," *Rapid Commun. Mass Spectrom.*, vol. 18, no. 18, pp. 2131–2145, Sep. 2004.

[109] P. Shipkova, D. M. Drexler, R. Langish, J. Smalley, M. E. Salyan, and M. Sanders, "Application of ion trap technology to liquid chromatography/mass spectrometry quantitation of large peptides," *Rapid Commun. Mass Spectrom.*, vol. 22, no. 9, pp. 1359–1366, May 2008.

[110] R. Kiyonami, A. Schoen, A. Prakash, S. Peterman, V. Zabrouskov, P. Picotti, R. Aebersold, A. Huhmer, and B. Domon, "Increased Selectivity, Analytical Precision, and Throughput in Targeted Proteomics," *Mol. Cell. Proteomics*, vol. 10, no. 2, p. M110.002931, Feb. 2011.

[111] T. A. Addona, S. E. Abbatiello, B. Schilling, S. J. Skates, D. R. Mani, D. M. Bunk, C. H. Spiegelman, L. J. Zimmerman, A.-J. L. Ham, H. Keshishian, S. C. Hall, S. Allen, R. K. Blackman, C. H. Borchers, C. Buck, H. L. Cardasis, M. P. Cusack, N. G. Dodder, B. W. Gibson, J. M. Held, T. Hiltke, A. Jackson, E. B. Johansen, C. R. Kinsinger, J. Li, M. Mesri, T. A. Neubert, R. K. Niles, T. C. Pulsipher, D. Ransohoff, H. Rodriguez, P. A. Rudnick, D. Smith, D. L. Tabb, T. J. Tegeler, A. M. Variyath, L. J. Vega-Montoto, Å. Wahlander, S. Waldemarson, M. Wang, J. R.

Whiteaker, L. Zhao, N. L. Anderson, S. J. Fisher, D. C. Liebler, A. G. Paulovich, F. E. Regnier, P. Tempst, and S. A. Carr, "Multi-site assessment of the precision and reproducibility of multiple reaction monitoring–based measurements of proteins in plasma," *Nat. Biotechnol.*, vol. 27, no. 7, pp. 633–641, Jul. 2009.

[112] I. Cima, R. Schiess, P. Wild, M. Kaelin, P. Schüffler, V. Lange, P. Picotti, R. Ossola, A. Templeton, O. Schubert, T. Fuchs, T. Leippold, S. Wyler, J. Zehetner, W. Jochum, J. Buhmann, T. Cerny, H. Moch, S. Gillessen, R. Aebersold, and W. Krek, "Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer," *Proc. Natl. Acad. Sci.*, vol. 108, no. 8, pp. 3342–3347, Feb. 2011.

[113] P. Picotti, B. Bodenmiller, L. N. Mueller, B. Domon, and R. Aebersold, "Full Dynamic Range Proteome Analysis of S. cerevisiae by Targeted Proteomics," *Cell*, vol. 138, no. 4, pp. 795–806, Aug. 2009.

[114] J. A. Loo, H. R. Udseth, R. D. Smith, and J. H. Futrell, "Collisional effects on the charge distribution of ions from large molecules, formed by electrospray-ionization mass spectrometry," *Rapid Commun. Mass Spectrom.*, vol. 2, no. 10, pp. 207–210, Dec. 1988.

[115] W. D. van Dongen, J. I. T. van Wijk, B. N. Green, W. Heerma, and J. Haverkamp, "Comparison between collision induced dissociation of electrosprayed protonated peptides in the up-front source region and in a low-energy collision cell," *Rapid Commun. Mass Spectrom.*, vol. 13, no. 17, pp. 1712–1716, Sep. 1999.

[116] E. R. Williams, S. Y. Loh, F. W. McLafferty, and R. B. Cody, "Hadamard transform measurement of tandem Fourier-transform mass spectra," *Anal. Chem.*, vol. 62, no. 7, pp. 698–703, Apr. 1990.

[117] J. Wilson and R. W. Vachet, "Multiplexed MS/MS in a Quadrupole Ion Trap Mass Spectrometer," *Anal. Chem.*, vol. 76, no. 24, pp. 7346–7353, Dec. 2004.

[118] J. D. Venable, M.-Q. Dong, J. Wohlschlegel, A. Dillin, and J. R. Yates, "Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra," *Nat. Methods*, vol. 1, no. 1, pp. 39–45, Oct. 2004.

[119] A. M. Graichen and R. W. Vachet, "Multiplexed MS/MS in a Miniature Rectilinear Ion Trap," *J. Am. Soc. Mass Spectrom.*, vol. 22, no. 4, pp. 683–688, Apr. 2011.

[120] A. Michalski, J. Cox, and M. Mann, "More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC–MS/MS," *J. Proteome Res.*, vol. 10, no. 4, pp. 1785–1793, Apr. 2011.

[121] E. J. Chang, V. Archambault, D. T. McLachlin, A. N. Krutchinsky, and B. T. Chait, "Analysis of Protein Phosphorylation by Hypothesis-Driven Multiple-Stage Mass Spectrometry," *Anal. Chem.*, vol. 76, no. 15, pp. 4472–4483, Aug. 2004.

[122] A. Panchaud, S. Jung, S. A. Shaffer, J. D. Aitchison, and D. R. Goodlett, "Faster, Quantitative, and Accurate Precursor Acquisition Independent From Ion Count," *Anal. Chem.*, vol. 83, no. 6, pp. 2250–2257, Mar. 2011.

[123] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold, "Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis," *Mol. Cell. Proteomics*, vol. 11, no. 6, p. O111.016717, Jun. 2012.

[124] T. Geiger, J. Cox, and M. Mann, "Proteomics on an Orbitrap Benchtop Mass Spectrometer Using All-ion Fragmentation," *Mol. Cell. Proteomics*, vol. 9, no. 10, pp. 2252–2261, Oct. 2010.

[125] J. D. Chapman, D. R. Goodlett, and C. D. Masselon, "Multiplexed and data-independent tandem mass spectrometry for global proteome profiling," *Mass Spectrom. Rev.*, p. n/a–n/a, 2013.

[126] C. A. Hastings, S. M. Norton, and S. Roy, "New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data," *Rapid Commun. Mass Spectrom.*, vol. 16, no. 5, pp. 462–467, Mar. 2002.

[127] M. Hilario, A. Kalousis, C. Pellegrini, and M. Müller, "Processing and classification of protein mass spectra," *Mass Spectrom. Rev.*, vol. 25, no. 3, pp. 409–449, May 2006.

[128] P. J. Statham, “Deconvolution and background subtraction by least-squares fitting with prefiltering of spectra,” *Anal. Chem.*, vol. 49, no. 14, pp. 2149–2154, décembre 1977.
[129] R. Gras, M. Müller, E. Gasteiger, S. Gay, P.-A. Binz, W. Bienvenut, C. Hoogland, J.-C. Sanchez, A. Bairoch, D. F. Hochstrasser, and R. D. Appel, “Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection,” *ELECTROPHORESIS*, vol. 20, no. 18, pp. 3535–3550, 1999.
[130] M. Gentzel, T. Köcher, S. Ponnusamy, and M. Wilm, “Preprocessing of tandem mass spectrometric data to support automatic protein identification,” *PROTEOMICS*, vol. 3, no. 8, pp. 1597–1610, Aug. 2003.
[131] A. Scherl, Y. S. Tsai, S. A. Shaffer, and D. R. Goodlett, “Increasing information from shotgun proteomic data by accounting for misassigned precursor ion masses,” *PROTEOMICS*, vol. 8, no. 14, pp. 2791–2797, 2008.
[132] K. R. Clauser, P. Baker, and A. L. Burlingame, “Role of Accurate Mass Measurement (±10 ppm) in Protein Identification Strategies Employing MS or MS/MS and Database Searching,” *Anal. Chem.*, vol. 71, no. 14, pp. 2871–2882, Jul. 1999.
[133] F. J. Dieguez-Acuna, S. A. Gerber, S. Kodama, J. E. Elias, S. A. Beausoleil, D. Faustman, and S. P. Gygi, “Characterization of Mouse Spleen Cells by Subtractive Proteomics,” *Mol. Cell. Proteomics*, vol. 4, no. 10, pp. 1459–1470, Oct. 2005.
[134] J. V. Olsen, S.-E. Ong, and M. Mann, “Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues,” *Mol. Cell. Proteomics*, vol. 3, no. 6, pp. 608–614, Jun. 2004.
[135] M. R. Hoopmann, G. L. Finney, and M. J. MacCoss, “High-Speed Data Reduction, Feature Detection, and MS/MS Spectrum Quality Assessment of Shotgun Proteomics Data Sets Using High-Resolution Mass Spectrometry,” *Anal. Chem.*, vol. 79, no. 15, pp. 5620–5632, Aug. 2007.
[136] S. Houel, R. Abernathy, K. Renganathan, K. Meyer-Arendt, N. G. Ahn, and W. M. Old, “Quantifying the Impact of Chimera MS/MS Spectra on Peptide Identification in Large-Scale Proteomics Studies,” *J. Proteome Res.*, vol. 9, no. 8, pp. 4152–4160, Aug. 2010.
[137] C. W. Ross, S. Guan, P. B. Grosshans, T. L. Ricca, and A. G. Marshall, “Two-dimensional Fourier transform ion cyclotron resonance mass spectrometry/mass spectrometry with stored-waveform ion radius modulation,” *J. Am. Chem. Soc.*, vol. 115, no. 17, pp. 7854–7861, Aug. 1993.
[138] P. Pfaendler, G. Bodenhausen, J. Rapin, M. E. Walser, and T. Gaeumann, “Broad-band two-dimensional Fourier transform ion cyclotron resonance,” *J. Am. Chem. Soc.*, vol. 110, no. 17, pp. 5625–5628, Aug. 1988.
[139] F. W. McLafferty, D. B. Stauffer, S. Y. Loh, and E. R. Williams, “Hadamard transform and ‘no-peak’ enhancement in measurement of tandem Fourier transform mass spectra,” *Anal. Chem.*, vol. 59, no. 17, pp. 2212–2213, Sep. 1987.
[140] C. Masselon, G. A. Anderson, R. Harkewicz, J. E. Bruce, L. Pasa-Tolic, and R. D. Smith, “Accurate Mass Multiplexed Tandem Mass Spectrometry for High-Throughput Polypeptide Identification from Mixtures,” *Anal. Chem.*, vol. 72, no. 8, pp. 1918–1924, Apr. 2000.
[141] T. P. Conrads, G. A. Anderson, T. D. Veenstra, L. Paša-Tolić, and R. D. Smith, “Utility of Accurate Mass Tags for Proteome-Wide Protein Identification,” *Anal. Chem.*, vol. 72, no. 14, pp. 3349–3354, juillet 2000.
[142] M. S. Lipton, L. Pasa-Tolic’, G. A. Anderson, D. J. Anderson, D. L. Auberry, J. R. Battista, M. J. Daly, J. Fredrickson, K. K. Hixson, H. Kostandarithes, C. Masselon, L. M. Markillie, R. J. Moore, M. F. Romine, Y. Shen, E. Stritmatter, N. Tolic’, H. R. Udseth, A. Venkateswaran, K.-K. Wong, R. Zhao, and R. D. Smith, “Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 17, pp. 11049–11054, Aug. 2002.
[143] L. Pasa-Tolić, C. Masselon, R. C. Barry, Y. Shen, and R. D. Smith, “Proteomic analyses using an accurate mass and time tag strategy,” *BioTechniques*, vol. 37, no. 4, pp. 621–624, 626–633, 636 passim, Oct. 2004.

[144] J. D. Egertson, A. Kuehn, G. E. Merrihew, N. W. Bateman, B. X. MacLean, Y. S. Ting, J. D. Canterbury, D. M. Marsh, M. Kellmann, V. Zabrouskov, C. C. Wu, and M. J. MacCoss, "Multiplexed MS/MS for improved data-independent acquisition," *Nat. Methods*, vol. 10, no. 8, pp. 744–746, août 2013.
[145] C. Lawson and R. Hanson, *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics, 1995.
[146] J. C. Silva, M. V. Gorenstein, G.-Z. Li, J. P. C. Vissers, and S. J. Geromanos, "Absolute Quantification of Proteins by LCMSE A Virtue of Parallel ms Acquisition," *Mol. Cell. Proteomics*, vol. 5, no. 1, pp. 144–156, Jan. 2006.
[147] K. Blackburn, F. Mbeunkui, S. K. Mitra, T. Mentzel, and M. B. Goshe, "Improving Protein and Proteome Coverage through Data-Independent Multiplexed Peptide Fragmentation," *J. Proteome Res.*, vol. 9, no. 7, pp. 3621–3637, Jul. 2010.
[148] M. J. MacCoss, C. C. Wu, H. Liu, R. Sadygov, and J. R. Yates, "A Correlation Algorithm for the Automated Quantitative Analysis of Shotgun Proteomics Data," *Anal. Chem.*, vol. 75, no. 24, pp. 6912–6921, Dec. 2003.
[149] S. M. Hengel, E. Murray, S. Langdon, L. Hayward, J. O'Donoghue, A. Panchaud, T. Hupp, and D. R. Goodlett, "Data-independent Proteomic Screen Identifies Novel Tamoxifen Agonist that Mediates Drug Resistance," *J. Proteome Res.*, vol. 10, no. 10, pp. 4567–4578, Oct. 2011.
[150] T. J. Griffin, H. Xie, S. Bandhakavi, J. Popko, A. Mohan, J. V. Carlis, and L. Higgins, "iTRAQ Reagent-Based Quantitative Proteomic Analysis on a Linear Ion Trap Mass Spectrometer," *J. Proteome Res.*, vol. 6, no. 11, pp. 4200–4209, Nov. 2007.
[151] M. Bantscheff, M. Boesche, D. Eberhard, T. Matthieson, G. Sweetman, and B. Kuster, "Robust and Sensitive iTRAQ Quantification on an LTQ Orbitrap Mass Spectrometer," *Mol. Cell. Proteomics*, vol. 7, no. 9, pp. 1702–1713, Sep. 2008.
[152] J. D. Venable, T. Xu, D. Cociorva, and J. R. Yates, "Cross-Correlation Algorithm for Calculation of Peptide Molecular Weight from Tandem Mass Spectra," *Anal. Chem.*, vol. 78, no. 6, pp. 1921–1929, Mar. 2006.
[153] P. C. Carvalho, X. Han, T. Xu, D. Cociorva, M. da G. Carvalho, V. C. Barbosa, and J. R. Yates, "XDIA: improving on the label-free data-independent analysis," *Bioinformatics*, vol. 26, no. 6, pp. 847–848, Mar. 2010.
[154] P. C. Carvalho, T. Xu, X. Han, D. Cociorva, V. C. Barbosa, and J. R. Yates, "YADA: a tool for taking the most out of high-resolution spectra," *Bioinformatics*, vol. 25, no. 20, pp. 2734–2736, Oct. 2009.
[155] C. Fu, L. Di, X. Han, C. Soderstrom, M. Snyder, M. D. Troutman, R. S. Obach, and H. Zhang, "Aldehyde Oxidase 1 (AOX1) in Human Liver Cytosols: Quantitative Characterization of AOX1 Expression Level and Activity Relationship," *Drug Metab. Dispos.*, vol. 41, no. 10, pp. 1797–1804, Oct. 2013.
[156] H. L. Röst, G. Rosenberger, P. Navarro, L. Gillet, S. M. Miladinović, O. T. Schubert, W. Wolski, B. C. Collins, J. Malmström, L. Malmström, and R. Aebersold, "OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data," *Nat. Biotechnol.*, vol. 32, no. 3, pp. 219–223, Mar. 2014.
[157] Y. Liu, R. Hüttenhain, B. Collins, and R. Aebersold, "Mass spectrometric protein maps for biomarker discovery and clinical research," *Expert Rev. Mol. Diagn.*, vol. 13, no. 8, pp. 811–825, Nov. 2013.
[158] J.-P. Lambert, G. Ivosev, A. L. Couzens, B. Larsen, M. Taipale, Z.-Y. Lin, Q. Zhong, S. Lindquist, M. Vidal, R. Aebersold, T. Pawson, R. Bonner, S. Tate, and A.-C. Gingras, "Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition," *Nat. Methods*, vol. 10, no. 12, pp. 1239–1245, décembre 2013.
[159] R. G. Sadygov, Z. Hao, and A. F. R. Huhmer, "Charger: Combination of Signal Processing and Statistical Learning Algorithms for Precursor Charge-State Determination from Electron-Transfer Dissociation Spectra," *Anal. Chem.*, vol. 80, no. 2, pp. 376–386, Jan. 2008.

[160] M. Bern, G. Finney, M. R. Hoopmann, G. Merrihew, M. J. Toth, and M. J. MacCoss, "Deconvolution of Mixture Spectra from Ion-Trap Data-Independent-Acquisition Tandem Mass Spectrometry," *Anal. Chem.*, vol. 82, no. 3, pp. 833–841, Feb. 2010.
[161] J. Cox and M. Mann, "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification," *Nat. Biotechnol.*, vol. 26, no. 12, pp. 1367–1372, Dec. 2008.
[162] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann, "Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment," *J. Proteome Res.*, vol. 10, no. 4, pp. 1794–1805, Apr. 2011.
[163] B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, B. Frewen, R. Kern, D. L. Tabb, D. C. Liebler, and M. J. MacCoss, "Skyline: an open source document editor for creating and analyzing targeted proteomics experiments," *Bioinformatics*, vol. 26, no. 7, pp. 966–968, Apr. 2010.
[164] G. L. Finney, A. R. Blackler, M. R. Hoopmann, J. D. Canterbury, C. C. Wu, and M. J. MacCoss, "Label-Free Comparative Analysis of Proteomics Mixtures Using Chromatographic Alignment of High-Resolution µLC–MS Data," *Anal. Chem.*, vol. 80, no. 4, pp. 961–971, Feb. 2008.
[165] C. Aubron, A. Nichol, D. J. Cooper, and R. Bellomo, "Age of red blood cells and transfusion in critically ill patients," *Ann. Intensive Care*, vol. 3, no. 1, p. 2, Jan. 2013.
[166] Marik PE and Sibbald WJ, "EFfect of stored-blood transfusion on oxygen delivery in patients with sepsis," *JAMA*, vol. 269, no. 23, pp. 3024–3029, juin 1993.
[167] P. C. Hébert, G. Wells, M. A. Blajchman, J. Marshall, C. Martin, G. Pagliarello, M. Tweeddale, I. Schweitzer, and E. Yetisir, "A Multicenter, Randomized, Controlled Clinical Trial of Transfusion Requirements in Critical Care," *N. Engl. J. Med.*, vol. 340, no. 6, pp. 409–417, février 1999.
[168] P. E. Marik and H. L. Corwin, "Efficacy of red blood cell transfusion in the critically ill: A systematic review of the literature*:," *Crit. Care Med.*, vol. 36, no. 9, pp. 2667–2674, Sep. 2008.
[169] I. Chin-Yee, N. Arya, and M. S. d' Almeida, "The Red Cell Storage Lesion and its Implication for Transfusion," *Transfus. Sci.*, vol. 18, no. 3, pp. 447–458, Sep. 1997.
[170] L. C. Wolfe, "The membrane and the lesions of storage in preserved red cells," *Transfusion (Paris)*, vol. 25, no. 3, pp. 185–203, mai 1985.
[171] R. T. Card, N. Mohandas, H. A. Perkins, and S. B. Shohet, "Deformability of stored red blood cells. Relationship to degree of packing," *Transfusion (Paris)*, vol. 22, no. 2, pp. 96–101, Mar. 1982.
[172] V. Pettilä, A. J. Westbrook, A. D. Nichol, M. J. Bailey, E. M. Wood, G. Syres, L. E. Phillips, A. Street, C. French, L. Murray, N. Orford, J. D. Santamaria, R. Bellomo, D. J. Cooper, "Age of red blood cells and mortality in the critically ill," *Crit. Care*, vol. 15, no. 2, p. R116, Apr. 2011.
[173] J. A. Weinberg, G. McGwin, M. J. Vandromme, M. B. Marques, S. M. Melton, D. A. Reiff, J. D. Kerby, and L. W. Rue, "Duration of Red Cell Storage Influences Mortality After Trauma," *J. Trauma*, vol. 69, no. 6, pp. 1427–1432, Dec. 2010.
[174] Z. Murrell, J. S. Haukoos, B. Putnam, and S. R. Klein, "The Effect of Older Blood on Mortality, Need for ICU Care, and the Length of ICU Stay After Major Trauma," *Am. Surg.*, vol. 71, no. 9, pp. 781–785, Sep. 2005.
[175] A. H. M. van Straten, M. A. Soliman Hamad, A. A. J. van Zundert, E. J. Martens, J. F. ter Woorst, A. M. de Wolf, and V. Scharnhorst, "Effect of duration of red blood cell storage on early and late mortality after coronary artery bypass grafting," *J. Thorac. Cardiovasc. Surg.*, vol. 141, no. 1, pp. 231–237, Jan. 2011.
[176] P. C. Hébert, I. Chin-Yee, D. Fergusson, M. Blajchman, R. Martineau, J. Clinch, and B. Olberg, "A pilot trial evaluating the clinical effects of prolonged storage of red cells," *Anesth. Analg.*, vol. 100, no. 5, pp. 1433–1438, table of contents, May 2005.
[177] R. W. Taylor, J. O'Brien, S. J. Trottier, L. Manganaro, M. Cytron, M. F. Lesko, K. Arnzen, C. Cappadoro, M. Fu, M. S. Plisco, F. G. Sadaka, and C. Veremakis, "Red blood cell transfusions

and nosocomial infections in critically ill patients," *Crit. Care Med.*, vol. 34, no. 9, pp. 2302–2308; quiz 2309, Sep. 2006.
[178] A. B. Zimrin and J. R. Hess, "Current issues relating to the transfusion of stored red blood cells," *Vox Sang.*, vol. 96, no. 2, pp. 93–103, février 2009.
[179] A. D'Alessandro, P. G. Righetti, and L. Zolla, "The Red Blood Cell Proteome and Interactome: An Update," *J. Proteome Res.*, vol. 9, no. 1, pp. 144–163, Jan. 2010.
[180] E. M. Pasini, M. Kirkegaard, P. Mortensen, H. U. Lutz, A. W. Thomas, and M. Mann, "In-depth analysis of the membrane and cytosolic proteome of red blood cells," *Blood*, vol. 108, no. 3, pp. 791–801, Aug. 2006.
[181] F. Roux-Dalvai, A. G. de Peredo, C. Simó, L. Guerrier, D. Bouyssié, A. Zanella, A. Citterio, O. Burlet-Schiltz, E. Boschetti, P. G. Righetti, and B. Monsarrat, "Extensive Analysis of the Cytoplasmic Proteome of Human Erythrocytes Using the Peptide Ligand Library Technology and Advanced Mass Spectrometry," *Mol. Cell. Proteomics*, vol. 7, no. 11, pp. 2254–2269, Nov. 2008.
[182] G. M. D'Amici, S. Rinalducci, and L. Zolla, "Proteomic Analysis of RBC Membrane Protein Degradation during Blood Storage," *J. Proteome Res.*, vol. 6, no. 8, pp. 3242–3255, août 2007.
[183] G. J. C. G. M. Bosman, E. Lasonder, M. Luten, B. Roerdinkholder-Stoelwinder, V. M. J. Novotný, H. Bos, and W. J. De Grip, "The proteome of red cell membranes and vesicles during storage in blood bank conditions," *Transfusion (Paris)*, vol. 48, no. 5, pp. 827–835, May 2008.
[184] D. Coelho Graça, P. Lescuyer, L. Clerici, Y. O. Tsybin, R. Hartmer, M. Meyer, K. Samii, D. F. Hochstrasser, and A. Scherl, "Electron transfer dissociation mass spectrometry of hemoglobin on clinical samples," *J. Am. Soc. Mass Spectrom.*, vol. 23, no. 10, pp. 1750–1756, Oct. 2012.
[185] M. Dakna, Z. He, W. C. Yu, H. Mischak, and W. Kolch, "Technical, bioinformatical and statistical aspects of liquid chromatography-mass spectrometry (LC-MS) and capillary electrophoresis-mass spectrometry (CE-MS) based clinical proteomics: a critical assessment," *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.*, vol. 877, no. 13, pp. 1250–1258, May 2009.
[186] F. Gluck, C. Hoogland, P. Antinori, X. Robin, F. Nikitin, A. Zufferey, C. Pasquarello, V. Fétaud, L. Dayon, M. Müller, F. Lisacek, L. Geiser, D. Hochstrasser, J.-C. Sanchez, and A. Scherl, "EasyProt--an easy-to-use graphical platform for proteomics data analysis," *J. Proteomics*, vol. 79, pp. 146–160, Feb. 2013.
[187] E. Ahrné, F. Nikitin, F. Lisacek, and M. Müller, "QuickMod: A Tool for Open Modification Spectrum Library Searches," *J. Proteome Res.*, vol. 10, no. 7, pp. 2913–2921, Jul. 2011.
[188] E. Ahrné, Y. Ohta, F. Nikitin, A. Scherl, F. Lisacek, and M. Müller, "An improved method for the construction of decoy peptide MS/MS spectra suitable for the accurate estimation of false discovery rates," *PROTEOMICS*, vol. 11, no. 20, pp. 4085–4095, 2011.
[189] C. G. DeVry and S. Clarke, "Polymorphic forms of the protein L-isoaspartate (D-aspartate) O-methyltransferase involved in the repair of age-damaged proteins," *J. Hum. Genet.*, vol. 44, no. 5, pp. 275–288, 1999.
[190] H. Wang, D. S. Kennedy, K. D. Nugent, G. K. Taylor, and D. R. Goodlett, "A Qit-q-Tof mass spectrometer for two-dimensional tandem mass spectrometry," *Rapid Commun. Mass Spectrom.*, vol. 21, no. 19, pp. 3223–3226, Oct. 2007.

8 Annexes

8.1 MS/MS clustering source code

core package:

```
import java.io.File;
import java.io.PrintWriter;
import java.text.ParseException;
import java.io.IOException;

import java.util.ArrayList;
import java.util.List;
import java.util.Set;
import java.util.TreeSet;

import org.expasy.jpl.core.ms.lc.RetentionTime;
import org.expasy.jpl.core.ms.spectrum.PeakList;
import org.expasy.jpl.core.ms.spectrum.PeakListImpl;
import org.expasy.jpl.core.ms.spectrum.peak.PeakImpl;
import org.expasy.jpl.io.ms.MassSpectrum;

import cern.colt.matrix.DoubleMatrix2D;
import cern.colt.matrix.impl.DenseDoubleMatrix2D;

import org.jfree.data.xy.XYDataset;
import org.jfree.data.xy.XYSeries;
import org.jfree.data.xy.XYSeriesCollection;

import tools.FilesFinder;
import tools.Tools;


import cern.colt.list.DoubleArrayList;


public class Reader_core {

	File file;
	Pattern scan_pattern;
	boolean preprocessing;
	List<MassSpectrum> uniqueMassSpectra;
	Set<Double> channels;
	List<List<MassSpectrum>> dataPerChannel;

	public Reader_core (File file){
		this.file = file;
		this.scan_pattern = Pattern.compile("[\\w+\\_]+\\.(\\d+).*?");
	}
	public Reader_core (File file, String pattern, boolean preprocessing){
		this.file = file;
		this.scan_pattern = Pattern.compile(pattern);
		this.preprocessing = preprocessing;
	}

	/**
	 *
	 * @param file
	 * @param preprocessing
```

```
	 */
	public Reader_core (File file, boolean preprocessing){//in use
		this.file = file;
		this.preprocessing = preprocessing;
		this.scan_pattern = Pattern.compile("[\\w+\\_]+\\.(\\d+).*?");
	}


	public void FilterUniqueSP() throws ParseException{//in use

		this.uniqueMassSpectra = new ArrayList<MassSpectrum>();
		this.channels = new TreeSet<Double>();

		MGFReader reader =  MGFReader.newInstance();
		reader.enableAutoScanNum(true);
		reader.parse(this.file);
		AbstractExtraIterator<MassSpectrum> it = reader.iterator();

		MassSpectrum spectrum;
		NPeakGroupsPerWindowFilter filter = new NPeakGroupsPerWindowFilter(4,
10, 1.5);
		PeakList filter_peakList;
		Matcher matcher;

		int nb_total = 0;
		int nb_unique = 0;

		while(it.hasNext()){
			spectrum = it.next();
			if(spectrum.getPeakList().hasIntensities() == true &&
spectrum.getPeakList().getPrecursor().getCharge() == 2){  //non-empty peaklist
				matcher =
this.scan_pattern.matcher(spectrum.getTitle());
				if(matcher.matches()){
					if(this.preprocessing == true){
						filter_peakList =
filter.transform(spectrum.getPeakList());
						spectrum.setPeakList(filter_peakList);

	spectrum.setScanNum(Integer.parseInt(matcher.group(1)));
						this.uniqueMassSpectra.add(spectrum);

	this.channels.add(spectrum.getPeakList().getPrecursor().getMz());

						nb_unique++;
					}
					else{

	spectrum.setScanNum(Integer.parseInt(matcher.group(1)));
						this.uniqueMassSpectra.add(spectrum);

	this.channels.add(spectrum.getPeakList().getPrecursor().getMz());
						nb_unique++;
					}
				}
			}
			nb_total++;
		}
```

```
			System.out.println("total spectra = "+nb_total);
			System.out.println("unique spectra = "+nb_unique);

		}


		/**
		 * - grouping MS/MS data per channel
		 * @param channel_tol
		 */
		public void MassSpectraPerChannel(double channel_tol){

			this.dataPerChannel = new ArrayList<List<MassSpectrum>>();
			for(double channel : this.channels){
				List<MassSpectrum> tmp = new ArrayList<MassSpectrum>();
				for(MassSpectrum spectrum : this.uniqueMassSpectra){
					if(Math.abs(channel-
spectrum.getPeakList().getPrecursor().getMz()) < channel_tol){
						tmp.add(spectrum);
					}
				}
				this.dataPerChannel.add(tmp);
			}
		}

		public List<MassSpectrum> getUniqueSP(){
			return this.uniqueMassSpectra;
		}
		public List<Double> getChannels(){
			List<Double> u_channels = new ArrayList<Double>();
			u_channels.addAll(this.channels);
			return u_channels;
		}
		public List<List<MassSpectrum>>getDataPerChannel(){
			return this.dataPerChannel;
		}
}


public class Binner_core {

		double min;
		double max;
		double increment;

		List<List<BinPeak>> metaBinData;
		DoubleArrayList bins;
		DoubleArrayList binRTs;
		DoubleMatrix2D channelMatrix;

		List<PeakList> binnedMatrixPL;

		/**
		 * bin MS/MS peaks
		 * @param min                   - min. val. of bin
		 * @param max                   - max. val. of bin
		 * @param increment       - bin size be carefull! The value is the total
size not the centered value
		 */
```

```
	public Binner_core(double min, double max, double increment){
		this.min = min;
		this.max = max;
		this.increment = increment;
	}

	// build bin data
	/**
	 * construct the MS/MS peak bin data
	 * @param msPerChannel
	 */
	public void BinList(List<MassSpectrum> msPerChannel){
		this.metaBinData = new ArrayList<List<BinPeak>>();
		this.bins = new DoubleArrayList();
		Set<Double> binRT = new TreeSet<Double>();

		for(double i = this.min; i <= this.max; i +=this.increment ){
			List<BinPeak> tmp = new ArrayList<BinPeak>();
			for(MassSpectrum spectrum : msPerChannel){
				for(int j = 0; j < spectrum.getPeakList().size(); j++){
					if(Math.abs(spectrum.getPeakList().getMzAt(j) - i)
< (this.increment/2)){
						tmp.add(new
BinPeak(spectrum.getPeakList().getMzAt(j),
spectrum.getPeakList().getIntensityAt(j),
spectrum.getRetentionTime().getValue()));
					}
				}
				binRT.add(spectrum.getRetentionTime().getValue());
			}

			this.metaBinData.add(tmp);
			this.bins.add(i);
		}

		this.binRTs = new DoubleArrayList();
		this.binRTs.addAllOf(binRT);

		this.bins.trimToSize();
		this.binRTs.trimToSize();
	}

	//convert bin data into matrix
	/**
	 * convert the binned MS/MS data into matrix, with row = nb. bin and colum =
retention. coincident peaks within the m/z and rt tolerance are summed for
intensity
	 * @param channel		- value of channel window
	 * @param binpl			- should binned MS/MS peaks list be generated?
	 * @throws IOException
	 */
	public void intoMatrix(double channel, boolean binpl) throws IOException{

		this.binnedMatrixPL = new ArrayList<PeakList>();

		this.channelMatrix = new
DenseDoubleMatrix2D(this.bins.size(),this.binRTs.size());
		this.channelMatrix.assign(0.0);
```

```
		for(int i = 0; i < this.bins.size(); i++){
			for(int j = 0; j < this.binRTs.size(); j++){
				for(BinPeak binpeak : this.metaBinData.get(i)){
					if(binpeak.getRT() == this.binRTs.get(j)){
						double r =
this.channelMatrix.getQuick(i,j);
						r = r+binpeak.getIntensity();
						this.channelMatrix.setQuick(i, j, r);
					}
				}
			}
		}

		this.channelMatrix.trimToSize();

		if(binpl){
			DoubleArrayList mzs;
			DoubleArrayList its;

			RetentionTime rt;
			PeakList pl;

			for(int i = 0; i < this.channelMatrix.columns(); i++){

				mzs = new DoubleArrayList();
				its = new DoubleArrayList();
				for(int j = 0; j < this.channelMatrix.rows(); j++){
					if(this.channelMatrix.getQuick(j, i) > 0){
						mzs.add(this.bins.getQuick(j));
						its.add(this.channelMatrix.getQuick(j, i));
					}
				}

				mzs.trimToSize();
				its.trimToSize();

				rt = new RetentionTime();
				rt.setValue(this.binRTs.getQuick(i));
				pl = new
PeakListImpl.Builder(mzs.elements()).intensities(its.elements()).
							precursor(new
PeakImpl.Builder(channel).charge(2).msLevel(2).rt(rt).build()).
							build();

				this.binnedMatrixPL.add(pl);
			}
		}

	}

	public List<PeakList> getBinnedMatrixPeakList(){
		return this.binnedMatrixPL;
	}
	public DoubleArrayList getBinMzs(){
		return this.bins;
	}
	public List<List<BinPeak>> getMetaBinData(){
		return this.metaBinData;
	}
```

```
	public DoubleArrayList getBinRTs(){
		return this.binRTs;
	}
	public DoubleMatrix2D getChannelMatrix(){
		return this.channelMatrix;
	}

}



public class BinPeak {

	double mz;
	double intensity;
	double retention;

	public BinPeak(double mz, double intensity, double retention){
		this.mz = mz;
		this.intensity = intensity;
		this.retention = retention;
	}

	public double getMz(){
		return mz;
	}

	public double getIntensity(){
		return intensity;
	}

	public double getRT(){
		return this.retention;
	}
}



public class LMData {

	double leftB;
	double maxB;
	double rightB;

	int left_idx;
	int max_idx;
	int right_idx;

	public LMData(double leftB, double maxB, double rightB){

		this.leftB = leftB;
		this.maxB = maxB;
		this.rightB = rightB;
	}

	public LMData(int left_idx, int max_idx, int right_idx){
		this.left_idx = left_idx;
		this.max_idx = max_idx;
		this.right_idx = right_idx;
```

```
	}

	public double getLeftBound(){
		return leftB;
	}
	public double getRightBound(){
		return rightB;
	}
	public double getMax(){
		return maxB;
	}

	public int getLeftIdx(){
		return left_idx;
	}
	public int getRightIdx(){
		return right_idx;
	}
	public int getMaxIdx(){
		return max_idx;
	}

	public void setLeftBound(double leftBound){
		leftB = leftBound;
	}
	public void setMax(double max){
		maxB = max;
	}
	public void setRighttBound(double rightBound){
		rightB = rightBound;
	}

	public void setLeftBound(int leftBound){
		left_idx = leftBound;
	}
	public void setMax(int max){
		max_idx = max;
	}
	public void setRighttBound(int rightBound){
		right_idx = rightBound;
	}
}


public class LocalMaxima_core {


	DoubleMatrix2D chrom_matrix;
	DoubleArrayList rt_mat;
	DoubleArrayList mz_mat;
	Double channel;
	List<List<LMData>> LMperBin;
	List<PeakList> consensus_S1;

	/**
	 * stage 1   - extract local maxima from XIC of MS/MS for a given bin m/z
value
	 * @param chrom_matrix
```

```
         * @param rt_mat
         * @param mz_mat
         * @param channel
         */
        LocalMaxima_core(DoubleMatrix2D chrom_matrix, DoubleArrayList rt_mat,
DoubleArrayList mz_mat, double channel){
                this.chrom_matrix = chrom_matrix;
                this.rt_mat = rt_mat;
                this.mz_mat = mz_mat;
                this.channel = channel;
        }

        /**
         * stage 1   - contains local maxima per XICs, per bin
         * @return
         */
        public List<List<LMData>> getLocalMaxPerBin(){
                return this.LMperBin;
        }

        /**
         * stgae 1   - consensus spectra
         * @return
         */
        public List<PeakList> getConsensus_S1(){
                return this.consensus_S1;
        }

        /**
         * stage 1
         * @param intensity_cut
         */
        public void findLocalMaximaPerBin(double intensity_cut){

                this.LMperBin = new ArrayList<List<LMData>>();
                double [] retention = this.rt_mat.elements();

                for(int k = 0; k < this.chrom_matrix.rows(); k++){
                        double [] intensity = this.chrom_matrix.viewRow(k).toArray();
                        localmaxima_v1(retention, intensity, intensity_cut);
                }

        }

        /**
         * stage 1 - latest version of local maxima
         * @param retention
         * @param intensity
         * @param intenisty_cut
         */
        private void localmaxima_v1(double[] retention, double[] intensity, double
intenisty_cut){

                List<LMData> maxima = new ArrayList<LMData>();
                // Find maxima
                for (int i = 0; i < retention.length; i++) {
                        int next = Math.min(i + 1, retention.length - 1);
                        int begin = i;
                        boolean isMax = false;
```

```
                    if(intensity[next] > intenisty_cut){
                          isMax = true;
                          while(intensity[next] > intenisty_cut && next <
retention.length-1){
                                i++;
                                next++;
                          }
                    }
                    if(isMax){
                          if(next <= retention.length){
                                int max = begin;
                                for(int j = begin; j < next; j++){
                                      if(intensity[j] > intensity[max]){
                                            max = j;
                                      }
                                }
                                //maxima.add(new LMData(retention[begin],
retention[max], retention[next]));
                                maxima.add(new LMData(begin, max, next));
                          }
                    }
              }
              this.LMperBin.add(maxima);
        }

        /**
         * stage 1
         * @param type
         * @param is_consensus
         * @throws IOException
         */
        public void clusterLocalMaximaOnly(String type, boolean is_consensus) throws
IOException{

              DoubleMatrix2D data = this.chrom_matrix;
              DoubleMatrix2D maxima_matrix = data.copy();
              maxima_matrix.assign(0.0);
              List<List<LMData>> lmPerBin = this.LMperBin;

              Set<Integer> uniqueMax = new TreeSet<Integer>();

              for(int i = 0; i < data.rows(); i++){
                    for(LMData lm : lmPerBin.get(i)){
                          maxima_matrix.setQuick(i, lm.getMaxIdx(), 1.0);
                          uniqueMax.add(lm.getMaxIdx());
                    }
              }

              if(is_consensus){
                    DoubleArrayList retention = this.rt_mat;
                    DoubleArrayList bins = this.mz_mat;
                    DoubleArrayList mzs;
                    DoubleArrayList its;

                    RetentionTime rt;
                    PeakList pl;
                    this.consensus_S1 = new ArrayList<PeakList>();
```

```
                        for(int idx : uniqueMax){
                                mzs = new DoubleArrayList();
                                its = new DoubleArrayList();
                                for(int i = 0; i < maxima_matrix.rows(); i++){
                                        if(type.equals("strict")){
                                                if(maxima_matrix.get(i, idx) == 1 &&
data.getQuick(i, idx) > 0.0 ){
                                                        mzs.add(bins.getQuick(i));
                                                        its.add(data.getQuick(i, idx));
                                                }
                                        }
                                        else{
                                                if(data.getQuick(i, idx) > 0.0){
                                                        mzs.add(bins.getQuick(i));
                                                        its.add(data.getQuick(i, idx));
                                                }
                                        }
                                }
                                mzs.trimToSize();
                                its.trimToSize();

                                rt = new RetentionTime();
                                rt.setValue(retention.getQuick(idx));
                                pl = new
PeakListImpl.Builder(mzs.elements()).intensities(its.elements()).
                                                precursor(new
PeakImpl.Builder(this.channel).charge(2).msLevel(2).rt(rt).build()).
                                                build();
                                this.consensus_S1.add(pl);
                        }
                }
        }


        /**************************************************
         *  STAGE II
         **************************************************/

        List<List<PeakList>> clusterPL;
        DoubleArrayList clusterBin;
        List<LMData> bounds;
        /**
         * stage 2
         * @return
         */
        public List<List<PeakList>> getClusterPeakList(){
                return this.clusterPL;
        }

        /**
         * stage 2
         * @return
         */
        public DoubleArrayList getClusterRTBin(){
                return this.clusterBin;
        }

        /**
         *
```

```
         * @return
         */
        public List<LMData> getClusterBounds(){
                return this.bounds;
        }

        /**
         * stage 2 - local local maxima are automatically extracted by the number of
co-eluting maxima.
         * @param window - window centered on a rt within elution bounds must be
found
         * @param type - "strict for select only peaks within the bounds
         * @param limitLM - number of minimum co-elution maxima for auto-detection
of molecular ion elution
         * @param isStrict strict for non-ovrelaping bounds
         * @throws IOException
         */
        public void clusterLocalMaximaWindowSorted(double window,String type, int
limitLM, boolean isStrict) throws IOException{

                //matrix
                DoubleMatrix2D data = this.chrom_matrix;
                DoubleMatrix2D cluster_matrix = data.copy();
                cluster_matrix.assign(0.0);

                /*********************************************
                 * auto-detection of the cluster of local maxima
                 *********************************************/
                //local maixma data per bin IC
                List<List<LMData>> lmPerBinLi = this.LMperBin;

                //matrix with bound codes
                for(int i = 0; i < data.rows(); i++){
                        for(LMData lm : lmPerBinLi.get(i)){
                                int idx = lm.getLeftIdx();
                                while(idx <= lm.getRightIdx()){
                                        cluster_matrix.setQuick(i, idx, 1.0);
                                        idx++;
                                }
                        }
                }

                //regroupment of co-eluting of maxima indices
                Map<Integer, List<LMData>> clusterLM = new TreeMap<Integer,
List<LMData>>();
                for(List<LMData> lmPerBin : lmPerBinLi){
                        for(LMData lm : lmPerBin){
                                if(clusterLM.containsKey(lm.getMaxIdx())){
                                        clusterLM.get(lm.getMaxIdx()).add(lm);
                                }
                                else{
                                        List<LMData> tmp = new ArrayList<LMData>();
                                        tmp.add(lm);
                                        clusterLM.put(lm.getMaxIdx(), tmp);
                                }
                        }
                }
                // list of indices that passed the criteria
                List<Integer> cluster_indices = new ArrayList<Integer>();
```

```
		for(Map.Entry<Integer, List<LMData>> m :clusterLM.entrySet()){
			if(m.getValue().size() >= limitLM){
				cluster_indices.add(m.getKey());
				/*for(LMData lmd : m.getValue()){
					System.out.println(lmd.getLeftIdx()+" - 
"+lmd.getMaxIdx()+" - "+lmd.getRightIdx());
				}
				System.out.println("::::::::::::::::::::");*/
			}
		}

		//look for the most frequent bounds for a given cluster of local 
maxima
		List<LMData> second_clusters = new ArrayList<LMData>();
		for(int index : cluster_indices){
			if(clusterLM.containsKey(index)){
				Map<Integer, Integer> left = new HashMap<Integer, 
Integer>();
				Map<Integer, Integer> right = new HashMap<Integer, 
Integer>();
				for(LMData l : clusterLM.get(index)){
					//left
					if(left.containsKey(l.getLeftIdx())){
						int count = left.get(l.getLeftIdx())+1;
						left.put(l.getLeftIdx(), count);
					}
					else{
						left.put(l.getLeftIdx(), 1);
					}
					//right
					if(right.containsKey(l.getRightIdx())){
						int count = right.get(l.getRightIdx())+1;
						right.put(l.getRightIdx(), count);
					}
					else{
						right.put(l.getRightIdx(), 1);
					}
				}
				//left
				int l_idx = 0;
				int l = 0;
				for(Map.Entry<Integer, Integer> en : left.entrySet()){
					if(en.getValue() > l){
						l = en.getValue();
						l_idx = en.getKey();
					}
				}
				//right
				int r_idx = 0;
				int r = 0;
				for(Map.Entry<Integer, Integer> en : right.entrySet()){
					if(en.getValue() > r){
						r = en.getValue();
						r_idx = en.getKey();
					}
				}
				second_clusters.add(new LMData(l_idx, index, r_idx));
				//second_clusters.add(new LMData(this.rt_mat.get(l_idx), 
this.rt_mat.get(index), this.rt_mat.get(r_idx)));
```

```
                }
            }

            /*
            //non-overaling selection
            List<LMData> strictLM = new ArrayList<LMData>();
            for(int i = 1; i < second_clusters.size(); i++){
                if(i < second_clusters.size()){
                    int k = i-1;
                    int strict_l = second_clusters.get(k).getRightIdx();

      while(Math.abs(this.rt_mat.get(second_clusters.get(i).getMaxIdx()) -
this.rt_mat.get(second_clusters.get(k).getMaxIdx())) < window/2){

      if(this.rt_mat.get(second_clusters.get(i).getRightIdx()) >
this.rt_mat.get(strict_l)){
                                    strict_l =
second_clusters.get(i).getRightIdx();
                                }
                                i++;
                            }
                            strictLM.add(new
LMData(second_clusters.get(k).getLeftIdx(), second_clusters.get(i-1).getMaxIdx(),
strict_l));
                    }
            }
            */


            List<LMData> strictLM = new ArrayList<LMData>();
            for(int i = 0; i < second_clusters.size(); i++){

                int left = second_clusters.get(i).getLeftIdx();
                int mid = second_clusters.get(i).getMaxIdx();
                int right = second_clusters.get(i).getRightIdx();

                while(this.rt_mat.get(second_clusters.get(i).getMaxIdx()) -
this.rt_mat.get(mid) < window/2 && i < second_clusters.size()-1){
                        if(this.rt_mat.get(second_clusters.get(i).getLeftIdx())
< left){
                            left = second_clusters.get(i).getLeftIdx();
                        }
                        if(this.rt_mat.get(second_clusters.get(i).getRightIdx())
> right){
                            right = second_clusters.get(i).getRightIdx();
                        }
                        i++;
                }
                strictLM.add(new LMData(left, mid, right));
            }

            //consnesus spectra building
            this.clusterPL = new ArrayList<List<PeakList>>();

            if(isStrict == true){
                this.bounds = new ArrayList<LMData>(strictLM);
            }
            else{
                this.bounds = new ArrayList<LMData>(second_clusters);
```

```
			}

			for(LMData b : this.bounds){
				PeakList pl;
				RetentionTime rt;

				DoubleArrayList mzs;
				DoubleArrayList its;
				List<PeakList> consensus_pl = new ArrayList<PeakList>();

				for(int i = b.getLeftIdx(); i <= b.getRightIdx(); i++){

					mzs = new DoubleArrayList();
					its = new DoubleArrayList();
					for(int j = 0; j < cluster_matrix.rows(); j++){
						if(type.equals("strict")){
							if(cluster_matrix.getQuick(j, i) == 1.0 &&
data.getQuick(j, i) > 0.0){
								mzs.add(this.mz_mat.getQuick(j));
								its.add(data.getQuick(j, i));
							}
						}
						else{
							if(data.getQuick(j, i) > 0.0){
								mzs.add(this.mz_mat.getQuick(j));
								its.add(data.getQuick(j, i));
							}
						}
					}

					mzs.trimToSize();
					its.trimToSize();

					if(mzs.size() > 1){
						rt = new RetentionTime();
						rt.setValue(this.rt_mat.getQuick(i));
						pl = new
PeakListImpl.Builder(mzs.elements()).intensities(its.elements()).
								precursor(new
PeakImpl.Builder(this.channel).charge(2).msLevel(2).rt(rt).build()).
								build();
						consensus_pl.add(pl);
					}
				}
				this.clusterPL.add(consensus_pl);
			}
		}

}



public class SpectralNetwork {

		List<List<PeakList>> clusterPLs;
		DoubleArrayList clusterRTs;
		double cut;
		String title;
```

```
	/**
	 * overload not used
	 * @param clusterPLs
	 * @param clusterRTs
	 * @param cut
	 */
	public SpectralNetwork(List<List<PeakList>> clusterPLs, DoubleArrayList
clusterRTs, double cut){
		this.clusterPLs = clusterPLs;
		this.clusterRTs = clusterRTs;
		this.cut = cut;
	}

	/**
	 * the used one
	 * @param clusterPLs
	 * @param cut
	 * @param title
	 */
	public SpectralNetwork(List<List<PeakList>> clusterPLs, double cut, String
title){
		this.clusterPLs = clusterPLs;
		this.cut = cut;
		this.title = title;
	}

	/**
	 *
	 * @param out                    - file printing object the write the
processed data
	 * @param channel         - channel of isolation window
	 * @param cluster_tol     - tolerance for MS/MS consensus spectrum building
(merging)
	 * @throws IOException
	 */
	public void doNetwork(PrintWriter out, double channel, double cluster_tol)
throws IOException{

		System.out.println(this.clusterPLs.size());
		//CLTools.error_test();
		for(int i = 0; i < this.clusterPLs.size(); i++){

			if(this.clusterPLs.get(i).size() > 1){
				System.out.println("network :
"+this.clusterPLs.get(i).size());
				for(PeakList pl : this.clusterPLs.get(i)){

	System.out.print(pl.getPrecursor().getRT().getValue()+" || ");
				}
				System.out.println();
				NetworkMatrix netScore = new NetworkMatrix(cluster_tol,
this.clusterPLs.get(i));
				List<List<Double>> matrix_score = new
ArrayList<List<Double>>(netScore.getScoreMatrix());
				compute_MST(this.clusterPLs.get(i), matrix_score,
channel, cluster_tol, out);
			}
			else{
				System.out.println("non-network");
```

```
                        if(this.clusterPLs.get(i).isEmpty() == false){

      if(this.clusterPLs.get(i).get(0).hasIntensities()){

      Tools.precursorWriter(this.clusterPLs.get(i).get(0), out, channel,
this.title);
                              }
                              else{
                                    System.out.println("empty peaklist");
                              }
                        }
                        else{
                              System.out.println("empty cluster list");
                        }
                  }
            }
      }

      /**
       *
       * @param cluster_pl      - set of spectra from one cluster
       * @param matrix_score    - matrix that contains dot product score
(correlation)
       * @param channel         - channel of isolation window
       * @param cluster_tol     - tolerance for MS/MS consensus spectrum building
(merging)
       * @param out                  - file printing object the write the
processed data
       */
      private void compute_MST(List<PeakList> cluster_pl, List<List<Double>>
matrix_score, double channel, double cluster_tol, PrintWriter out){

            List<Double> rts = new ArrayList<Double>();
            for(PeakList pl : cluster_pl){
                  //System.out.println(pl.getPrecursor().getRT().getValue());
                  rts.add(pl.getPrecursor().getRT().getValue());
            }

            //build network
            NetworkDoGraph newGraph;
                        newGraph = new NetworkDoGraph(matrix_score, rts,
cluster_pl);

            //find the shortest path
            Graph<MyNode, MyLink> mstGraph;
                        mstGraph = newGraph.getMSTGraph();

            //save edges
            List<MyLink> edgeLi;
                        edgeLi = new ArrayList<MyLink>(mstGraph.getEdges());

            //remove links under threshold
            for(MyLink iLink : edgeLi){
                  if(iLink.getWeight() > this.cut){
                        mstGraph.removeEdge(iLink);
                  }
            }

            //group the nods according to the separation
```

```
            WeakComponentClusterer<MyNode, MyLink> wpcluster = new
WeakComponentClusterer<MyNode, MyLink>();
            Set<Set<MyNode>> cutcluster = new
HashSet<Set<MyNode>>(wpcluster.transform(mstGraph));

            List<PeakList> tmpCL;
            List<Double> tmpScan;
            double first;
            double last;

            //write grouped spectra after the filter
            for(Set<MyNode> iSet : cutcluster){

                  tmpCL = new ArrayList<PeakList>();
                  tmpScan = new ArrayList<Double>();

                  for(MyNode node_i : iSet){
                        tmpCL.add(node_i.getPeakList());
                        tmpScan.add(node_i.getScan());
                  }

                  Collections.sort(tmpScan);
                  first = tmpScan.get(0);
                  last = tmpScan.get(tmpScan.size()-1);

                  if(tmpScan.size() == 1){
                        Tools.Writer(tmpCL.get(0), out, channel, first,
this.title);
                  }

                  if(tmpScan.size() > 1){
                        SpectrumCluster singleSP =  new
SpectrumCluster.Builder().addSpectra(tmpCL).tol(cluster_tol).build();
                        singleSP.filter(tmpCL.size());
                        Tools.Writer(singleSP.getConsensusSpectrum(), out,
channel, first, last, this.title);
                  }

            }
      }


}


package network;

public class MyLink {

      double weight;
      String edge;

      /**
       *
       * @param weight (distance)
       * @param edge (symbol for connection)
       */
      public MyLink(double weight, String edge){
```

```
		this.weight = weight;
		this.edge = edge;
	}

	/**
	 * return the distance computed between a pair of nodes
	 * @return a double
	 */
	public double getWeight(){
		return this.weight;
	}

	/**
	 *
	 * @return a string as symbol of connection
	 */
	public String getEdge(){
		return this.edge;
	}

	/**
	 * return a string for graph legend
	 */
	public String toString(){
		return String.format("%.2f", weight);
	}
}

package network;

import org.expasy.jpl.core.ms.spectrum.PeakList;

/**
 * customed nodes
 * @author hup
 *
 */
public class MyNode {

	double scan;
	PeakList peakL;

	/**
	 *
	 * @param iScan
	 * @param peakL
	 */
	public MyNode(double scan, PeakList peakL){

		this.scan = scan;
		this.peakL = peakL;
	}

	/**
	 * return a scan that corresponds to a spectrum
	 * @return a double
	 */
	public double getScan(){
		return this.scan;
```

```
	}

	/**
	 * return a string
	 */
	public String toString(){
		return "node "+scan;
	}

	/**
	 * return spectrum that corresponds to a gievn scan
	 * @return a PeakList object
	 */
	public PeakList getPeakList(){

		return this.peakL;
	}
}

package network;

import java.util.ArrayList;
import java.util.List;

import org.apache.commons.collections15.Factory;
import org.apache.commons.collections15.Transformer;
import org.expasy.jpl.core.ms.spectrum.PeakList;

import edu.uci.ics.jung.algorithms.shortestpath.PrimMinimumSpanningTree;
import edu.uci.ics.jung.graph.Graph;
import edu.uci.ics.jung.graph.UndirectedSparseGraph;

/**
 * initiate graph with
 * @author hup
 *
 */
public class NetworkDoGraph {

	List<List<Double>> matrix;
	List<Double> scanLi;
	List<PeakList> clusterPL;

	/**
	 *
	 * @param matrix
	 * @param scanLi
	 * @param clusterPL
	 */
	public NetworkDoGraph(List<List<Double>> matrix, List<Double> scanLi,
			List<PeakList> clusterPL){
		this.matrix = matrix;
		this.scanLi = scanLi;
		this.clusterPL = clusterPL;
	}
```

```
	/**
	 * get a simple graph
	 * @return
	 */
	public Graph<MyNode, MyLink> getGraph(){
		//JUNG : generating graph and network for MST
		Factory<UndirectedSparseGraph<MyNode, MyLink>> graphFactory;

		// transformer
		Transformer<MyLink, Double> transF = new Transformer<MyLink, Double>(){
			public Double transform(MyLink link){
				return link.getWeight();
			}
		};

		//factory
		graphFactory = new Factory<UndirectedSparseGraph<MyNode, MyLink>>() {
			public UndirectedSparseGraph<MyNode, MyLink> create() {
				return new UndirectedSparseGraph<MyNode, MyLink>();
				}
		};

		// a new undirected grtaph is generated for each index i that corresponds to
cluster i
		Graph<MyNode, MyLink> g = new UndirectedSparseGraph<MyNode,
MyLink>();// undirected graph
		List<MyNode> nodLi = new ArrayList<MyNode>(); //nod are stored individually
a s index : nod(scan, spectrum) as JPLIMSPeakList or List<List<Double>>


		//initializing nods for cluster i
		for(int i = 0; i < scanLi.size(); i++){	//list of nods for i cluster
			nodLi.add(new MyNode(scanLi.get(i), clusterPL.get(i)));
		}

		//setting undirected graph for cluster i : generate a matrix based on SPC scoring
model
		for(int i = 0; i < scanLi.size(); i++){	//for each similarity matrix of i cluster do a
NetGraph
			for(int j = 0; j < scanLi.size(); j++){
					g.addEdge(new MyLink(
							1-(matrix.get(i).get(j)/matrix.get(i).get(i)),
							scanLi.get(i)+"->"+scanLi.get(j)),
nodLi.get(i), nodLi.get(j));
			}
		}

		return g;
	}

	/**
	 * Get graph based on the minimum spanning tree
	 * @return
```

```
	 */
	public Graph<MyNode, MyLink> getMSTGraph(){

		//JUNG : generating graph and network for MST algorithm
		Factory<UndirectedSparseGraph<MyNode, MyLink>> graphFactory;

		// mandatory method to use JUNG for network/graph generation ==> classe anonyme
		Transformer<MyLink, Double> transF = new Transformer<MyLink, Double>(){
			public Double transform(MyLink link){
				return link.getWeight();
			}
		};

		//==> classe anonyme
		graphFactory = new Factory<UndirectedSparseGraph<MyNode, MyLink>>() {
			public UndirectedSparseGraph<MyNode, MyLink> create() {
				return new UndirectedSparseGraph<MyNode, MyLink>();
				}
		};

		// a new undirected graph is generated for each index i that corresponds to cluster i
		Graph<MyNode, MyLink> g = new UndirectedSparseGraph<MyNode, MyLink>();// undirected graph
		List<MyNode> nodLi = new ArrayList<MyNode>(); //nod are stored individually a s index : nod(scan, spectrum) as JPLIMSPeakList or List<List<Double>>


		//initializing nods for cluster i
		for(int i = 0; i < scanLi.size(); i++){	//list of nods for i cluster
			nodLi.add(new MyNode(scanLi.get(i), clusterPL.get(i)));
		}

		//setting undirected graph for cluster i : generate a sinilarity matrix based on SPC scoring model
		for(int i = 0; i < scanLi.size(); i++){	//for each similarity matrix of i cluster do a NetGraph
			for(int j = 0; j < scanLi.size(); j++){
				g.addEdge(new MyLink(
						1-(matrix.get(i).get(j)/matrix.get(i).get(i)),
						scanLi.get(i)+"->"+scanLi.get(j)), nodLi.get(i), nodLi.get(j));
			}
		}

		PrimMinimumSpanningTree<MyNode,MyLink> mst = new PrimMinimumSpanningTree<MyNode, MyLink>(graphFactory,transF);
		Graph<MyNode,MyLink> mstGraph = mst.transform(g);	// new graph with the shortest pathways


		return mstGraph;
	}
```

```
}

package network;

import java.io.IOException;
import java.util.ArrayList;
import java.util.List;

import org.expasy.jpl.core.ms.spectrum.PeakList;
import org.expasy.jpl.msmatch.PeakListMatcherImpl;
import org.expasy.jpl.msmatch.model.PeakListBiGraphAlgoModel;
import org.expasy.jpl.msmatch.scorer.NCorrScorer;


public class NetworkMatrix {

	double peakMatchTol;
	List<PeakList> clusterPL;

	/**
	 * compute the matrix of peak dot product
	 * @param peakMatchTol - tolerance for peak correaltion dot product
	 * @param clusterPL - from a set of spectra from one cluster
	 */
	public NetworkMatrix(double peakMatchTol, List<PeakList> clusterPL){

		this.peakMatchTol = peakMatchTol;
		this.clusterPL = clusterPL;
	}

	/**
	 * return a matrix in List<List<Double>> with dot product for each element i, j
	 * @return a matrix of score based on correlation and relative intensity product
	 * @throws IOException
	 */
	public List<List<Double>> getScoreMatrix() throws IOException{

		List<List<Double>> score = new ArrayList<List<Double>>();
		PeakListMatcherImpl matcher;

		matcher = PeakListMatcherImpl.withTol(peakMatchTol);
		matcher.setScorer(NCorrScorer.getInstance());
		matcher.setAlgoModel(PeakListBiGraphAlgoModel
		   .newInstance(PeakListBiGraphAlgoModel.RELATIVE_INTENSITY_PROD_INV));

		for(int i = 0; i < clusterPL.size(); i++){
			List<Double> iScoreL = new ArrayList<Double>();
			for(int j = 0; j < clusterPL.size(); j++){
				matcher.computeMatch(clusterPL.get(i), clusterPL.get(j));
				iScoreL.add(matcher.getScore());
			}
			score.add(iScoreL);
		}
```

```
		return score;
	}

	/**
	 * return the setted tolerance for spectra matching
	 * @return double
	 */
	public double getPeakMatchTol(){
		return peakMatchTol;
	}

	/**
	 * return a list of spectra sent as parameters
	 * @return
	 */
	public List<PeakList> getClusterPL(){
		return clusterPL;
	}

}

package tools;
import java.io.File;

	public class FilesFinder {

		public FilesFinder() {
			super();
		}

		public  File[] findFiles(String directoryPath) {

			File directory = new File(directoryPath);
			File[] subfiles = null;

			if(!directory.exists()){

				System.out.println("Le fichier/rï¿½pertoire
'"+directoryPath+"' n'existe pas");

			}else if(!directory.isDirectory()){

				System.out.println("Le chemin '"+directoryPath+"'
correspond ï¿½ un fichier et non ï¿½ un rï¿½pertoire");

			}else{

				subfiles = directory.listFiles();
				String message = "Le rï¿½pertoire '"+directoryPath+"'
contient "+ subfiles.length+" fichier"+(subfiles.length>1?"s":"");
				System.out.println(message);


			}
			return subfiles;
		}
}
```

```
package tools;

import java.io.PrintWriter;

import org.expasy.jpl.core.ms.spectrum.PeakList;

public class Tools {

	/**
	 * generate MGF type feature
	 * @param pl
	 * @param out
	 * @param channel
	 * @param first
	 * @param name
	 */
	public static void Writer(PeakList pl, PrintWriter out, double channel, double first, String
name){

		int[] charges = {2,3};
		for(int charge : charges){

			out.println("BEGIN IONS");

	out.println("TITLE=CL_S1_S2_"+name+"."+(int)first+"."+(int)first+"."+charge);
			out.println("CHARGE="+charge+"+");
			out.println("PEPMASS="+channel);
			out.println("RTINSECONDS="+first);

			for(int j = 0; j < pl.size(); j++){

	//out.println(String.format("%.2f",pl.getMzs()[j])+"\t"+(int)pl.getIntensities()[j]);

	out.println(String.format("%.3f",pl.getMzs()[j])+"\t"+String.format("%.3f",pl.getIntensiti
es()[j]));
			}

			out.println();
			out.println("END IONS");
			out.println();
		}

	}

	/**
	 * generate MGF type feature
	 * @param pl
	 * @param out
	 * @param channel
	 * @param first
	 * @param last
	 * @param name
	 */
```

```
	public static void Writer(PeakList pl, PrintWriter out, double channel, double first,
double last, String name){

		int[] charges = {2,3};
		for(int charge : charges){

			out.println("BEGIN IONS");

	out.println("TITLE=CL_S1_S2_"+name+"."+(int)first+"."+(int)last+"."+charge);
			out.println("CHARGE="+charge+"+");
			out.println("PEPMASS="+channel);
			out.println("RTINSECONDS="+first);
			out.println("RTINSECONDS_CLUSTER="+first+";"+last);

			for(int j = 0; j < pl.size(); j++){

	//out.println(String.format("%.2f",pl.getMzs()[j])+"\t"+(int)pl.getIntensities()[j]);

	out.println(String.format("%.3f",pl.getMzs()[j])+"\t"+String.format("%.3f",pl.getIntensiti
es()[j]));
			}

			out.println();
			out.println("END IONS");
			out.println();
		}
	}

	/**
	 * generate MGF type feature for doomed data
	 * @param pl
	 * @param out
	 * @param count
	 * @param title
	 */
	public static void doomWriter(PeakList pl, PrintWriter out, int count, String title){

		int[] charges = {2,3};
		for(int charge : charges){

			out.println("BEGIN IONS");
			out.println("TITLE="+title+"."+count+"."+count+"."+charge);
			out.println("CHARGE="+charge+"+");
			out.println("PEPMASS="+pl.getPrecursor().getMz());
			out.println("RTINSECONDS="+pl.getPrecursor().getRT().getValue());

			for(int j = 0; j < pl.size(); j++){

	//out.println(String.format("%.2f",pl.getMzs()[j])+"\t"+(int)pl.getIntensities()[j]);

	out.println(String.format("%.3f",pl.getMzs()[j])+"\t"+String.format("%.3f",pl.getIntensiti
es()[j]));
			}
```

```
                        out.println();
                        out.println("END IONS");
                        out.println();
                }
        }

        public static void precursorWriter(PeakList pl, PrintWriter out, double precursor, String
title){

                //int[] charges = {2,3};
                //for(int charge : charges){

                        out.println("BEGIN IONS");

        out.println("TITLE="+title+"."+(int)pl.getPrecursor().getRT().getValue()+"."+(int)pl.getP
recursor().getRT().getValue()+"."+pl.getPrecursor().getCharge());
                        out.println("CHARGE="+pl.getPrecursor().getCharge()+"+");
                        out.println("PEPMASS="+precursor);
                        out.println("RTINSECONDS="+pl.getPrecursor().getRT().getValue());
                        for(int j = 0; j < pl.size(); j++){

        out.println(String.format("%.3f",pl.getMzs()[j])+"\t"+String.format("%.3f",pl.getIntensiti
es()[j]));
                        }
                        out.println();
                        out.println("END IONS");
                        out.println();
                //}
        }
}
```

8.2 Precursor ion m/z recalculation source code

```
package precursor;

import java.io.File;
import java.io.IOException;
import java.io.PrintWriter;
import java.text.ParseException;
import java.util.ArrayList;
import java.util.List;

import org.expasy.jpl.commons.collection.ExtraIterable.AbstractExtraIterator;
import org.expasy.jpl.core.ms.lc.RetentionTime;
import org.expasy.jpl.core.ms.spectrum.BinnedPeakList;
import org.expasy.jpl.core.ms.spectrum.BinnedPeakListImpl;
import org.expasy.jpl.core.ms.spectrum.PeakList;
import org.expasy.jpl.core.ms.spectrum.PeakListImpl;
import org.expasy.jpl.core.ms.spectrum.PeakListUtils;
import org.expasy.jpl.core.ms.spectrum.peak.PeakImpl;
import org.expasy.jpl.io.ms.MassSpectrum;
import org.expasy.jpl.io.ms.reader.MGFReader;
```

```
import tools.Tools;

import cern.colt.list.DoubleArrayList;
import cern.jet.stat.Descriptive;

import core.BinPeak;
import core.Binner_core;
import core.CLTools;

public class Read_precursor {


	File file;
	PeakList binPeakList;
	int [] sortedIntensityIndexDown;
	PrintWriter out;

	public Read_precursor(File file, PrintWriter out){
		this.file = file;
		this.out = out;
	}

	public void readMGF() throws ParseException, IOException{

		MGFReader reader =  MGFReader.newInstance();
		reader.enableAutoScanNum(true);
		reader.parse(this.file);
		AbstractExtraIterator<MassSpectrum> it = reader.iterator();

		MassSpectrum spectrum;
		while(it.hasNext()){
			spectrum = it.next();
			if(spectrum.getPeakList().hasIntensities()){
				processSpectrum(spectrum);

			}
		}
	}

	private void processSpectrum(MassSpectrum spectrum) throws IOException{
		//binning MS/MS spectrum
		double firstMz = spectrum.getPeakList().getMzAt(0);
		double lastMz =
spectrum.getPeakList().getMzAt(spectrum.getPeakList().size()-1);
		// bin class
		Binner_core binPL = new Binner_core(firstMz, lastMz, 2.0);
		List<MassSpectrum> binSpectrum = new ArrayList<MassSpectrum>();

		binSpectrum.add(spectrum);
		binPL.BinList(binSpectrum);     // wierd but the methods accepts only
a list of mass spectrum object

		List<List<BinPeak>> metaBinData = binPL.getMetaBinData();
		DoubleArrayList binMzs = binPL.getBinMzs();
		DoubleArrayList binmass = new DoubleArrayList();
		binMzs.trimToSize();
		DoubleArrayList binIntensities = new DoubleArrayList();
```

```
			// MS/MS bin spectrum processing : mean, median and more statistics
per bin m/z
			if(metaBinData.size() == binMzs.size()){
				for(int i = 0; i < binMzs.size(); i++){
					DoubleArrayList binIt = new DoubleArrayList();
					for(BinPeak peak : metaBinData.get(i)){
						binIt.add(peak.getIntensity());
					}
					binIt.trimToSize();
					if(Descriptive.sum(binIt) > 0){
						binIntensities.add(Descriptive.sum(binIt));  // or
whatever
						binmass.add(binMzs.getQuick(i));
					}
					//binIntensities.add(Descriptive.sum(binIt));// or
whatever
				}
			}

			binmass.trimToSize();
			binIntensities.trimToSize();

			// rebuilding peaklist
			RetentionTime rt = new RetentionTime();
			rt.setValue(spectrum.getRetentionTime().getValue());
			PeakList binpl = new
PeakListImpl.Builder(binmass.elements()).intensities(binIntensities.elements()).
					precursor(new
PeakImpl.Builder(spectrum.getPeakList().getPrecursor().getMz()).charge(spectrum.ge
tPeakList().getPrecursor().getCharge()).msLevel(2).rt(rt).build()).
					build();

			PeakListUtils tools = PeakListUtils.getInstance();
			this.sortedIntensityIndexDown =
tools.getSortedIndexIntensityDown(binpl);
			this.binPeakList = binpl;

			//compute precursor
			ComputePrecursor precursor = new ComputePrecursor(this.binPeakList,
this.sortedIntensityIndexDown);
			precursor.process(0.0, 3.0, this.file.getName().split("\\.")[0],
this.out);
		}

}



package precursor;

/**
 * store y and b ions, compute precursor ions based on complementarity, and,
cross-correlation score
 * @author hup
 *
 */
public class ComplementaryIons {

	double ionB;
```

```
	double ionY;
	double delta;
	double ionBIt;
	double ionYIt;
	int charge;

	/**
	 *
	 * @param ionB
	 * @param ionY
	 * @param delta
	 */
	public ComplementaryIons(double ionB, double ionY, double delta, int
charge){
		this.ionB = ionB;
		this.ionY = ionY;
		this.delta = delta;
		this.charge = charge;
	}

	/**
	 * @overload
	 * @param ionB
	 * @param ionY
	 * @param delta
	 * @param ionBIt
	 * @param ionYIt
	 * @param charge
	 */
	public ComplementaryIons(double ionB, double ionY, double delta, double
ionBIt, double ionYIt, int charge){
		this.ionB = ionB;
		this.ionY = ionY;
		this.delta = delta;
		this.ionBIt = ionBIt;
		this.ionYIt = ionYIt;
		this.charge = charge;
	}

	/**
	 *  get b ion
	 * @return double
	 */
	public double getIonB(){
		return ionB;
	}

	/**
	 * get y ion
	 * @return double
	 */
	public double getIonY(){
		return ionY;
	}

	/**
	 * get the difference between theo m/z and computed m/z
	 * @return double
	 */
```

```
	public double getDelta(){
		return delta;
	}

	/**
	 * compute precursor ion m/z based on y and b ion complementarity
	 * @return double
	 */
	public double getPrecursorX(){
		double precursor = 0;

		//precursor = ionB+ionY+delta;
		precursor = ionB+ionY;

		precursor = precursor/charge;
		return precursor;
	}

	/**
	 * compute cross score
	 * @return double
	 */
	public double getCrossScore(){
		double crossScore = ionYIt*(ionB+ionY);
		return crossScore;
	}

	public double getCrossIt(){
		double crossIt = ionYIt*(ionBIt);
		return crossIt;
	}

}
```

```
package precursor;


import graphic.BarGraph;

import graphic.CombinedGraph;


import java.io.IOException;

import java.io.PrintWriter;

import java.util.ArrayList;

import java.util.List;


import org.expasy.jpl.core.ms.spectrum.PeakList;

import org.expasy.jpl.io.ms.MassSpectrum;
```

```
import org.jfree.data.statistics.HistogramDataset;
import org.jfree.data.xy.XYDataset;
import org.jfree.ui.RefineryUtilities;

import tools.Tools;

import core.BinPeak;
import core.CLTools;

import cern.colt.list.DoubleArrayList;



public class ComputePrecursor {

	PeakList binPeakList;
	int [] sortedIntensityIndexDown;

	public ComputePrecursor(PeakList binPeakList, int [] sortedIntensityIndexDown){
		this.binPeakList =  binPeakList;
		this.sortedIntensityIndexDown = sortedIntensityIndexDown;
	}

	public void process(double intensity_cut, double vicinity_window, String title,
PrintWriter out) throws IOException{

		double precursorChannel = this.binPeakList.getPrecursor().getMz();
		int nPeaks = this.binPeakList.size();
		int charge = this.binPeakList.getPrecursor().getCharge();
```

```
double [] mzs = this.binPeakList.getMzs();
double [] its = this.binPeakList.getIntensities();
int t = mzs.length;
DoubleArrayList delta_Y = new DoubleArrayList();
List<ComplementaryIons> compIons = new ArrayList<ComplementaryIons>();

/**
 * @ for doubly charged peptide precursor ion
 */
if(charge == 2){
	for(int j = 0; j < nPeaks; j++ ){
		int iMin = t-1;
		int iMax = t-1;

		if(mzs[this.sortedIntensityIndexDown[j]] > precursorChannel &&
mzs[this.sortedIntensityIndexDown[j]] < precursorChannel*charge &&
its[this.sortedIntensityIndexDown[j]] > intensity_cut){
			double mzY = precursorChannel*charge-
mzs[this.sortedIntensityIndexDown[j]];
			if(mzY < precursorChannel && mzY >
this.binPeakList.getMzAt(0)){
				if(mzY < mzs[iMin]){
					while (mzs[iMin] > mzY - vicinity_window
&& iMin != 0){
						iMin--;
					}
					if(mzs[iMin] < mzY - vicinity_window){
						iMin = iMin+1;
					}
					while (mzs[iMax] > mzY + vicinity_window
&& iMax != 0){
```

```
                                        iMax--;
                                    }
                                    if(iMax-1 > iMin){
                                        for(int k = iMin; k < iMax; k++){
                                            delta_Y.add(mzs[k]-mzY);
//exp - th = delta y
                                            compIons.add(new
ComplementaryIons(mzs[this.sortedIntensityIndexDown[j]], mzs[k],mzs[k]-
mzY,its[this.sortedIntensityIndexDown[j]], its[k], charge));
                                        }
                                    }
                                }
                            }
                        }
                    }

                    System.out.println(delta_Y.size());

                    if(delta_Y.size() > 0){
                        delta_Y.trimToSize();
                        delta_Y.sort();

                        double minpre = delta_Y.get(0);
                        double maxpre = delta_Y.get(delta_Y.size()-1);
                        double step = 0.6;

                        List<List<Double>> occurence = new ArrayList<List<Double>>();
```

```
List<Double> occurence_indices = new ArrayList<Double>();


for(double i = minpre; i <= maxpre; i += step ){
        List<Double> tmp = new ArrayList<Double>();
        for(int j  = 0; j < delta_Y.size(); j++){
                if(Math.abs(delta_Y.getQuick(j) - i) < (step/2)){
                        tmp.add(delta_Y.getQuick(j));
                }
        }
        if(tmp.size() > 0){
                occurence_indices.add(i);
                occurence.add(tmp);
        }
}


int nb_occ = 3;
for(int i = 0; i < occurence.size(); i++){
        if(occurence.get(i).size() >= nb_occ){


                System.out.println("mass of precursor =
"+(this.binPeakList.getPrecursor().getMz()+occurence_indices.get(i)));
                Tools.precursorWriter(this.binPeakList, out,
(this.binPeakList.getPrecursor().getMz()+occurence_indices.get(i)), title);


        }
}



/*
```

```
                    HistogramDataset histogram = new HistogramDataset();
                    histogram.addSeries("test", delta_Y.elements(),
(int)Math.pow(delta_Y.size(),0.5));

                    BarGraph filterDistrib = new BarGraph("QM - Label-free Integration",
histogram);
                    filterDistrib.pack();
                    RefineryUtilities.centerFrameOnScreen(filterDistrib);
                    filterDistrib.setVisible(true);
                    */
                    //CLTools.error_test();

                }

        }
}
```