POLITECNICO DI MILANO

Scuola di Ingegneria dei Sistemi
Corso di Studi in Ingegneria Biomedica

IMPROVING QUANTIFICATION OF LABELED PEPTIDES IN MASS SPECTROMETRY-BASED PROTEOMICS

Student: Andrea Alamia - 751777

Supervisor: Prof. Linda Pattini
PhD. Salvatore Cappadona

Anno Accademico 2010-2011

Indice

Abstract

Bioinformatics is playing an increasingly important role in the field of biomedical sciences in the interpretation of the data in order to understand the mechanisms at the basis of the cellular functioning. In particular, proteomics is gaining more and more importance in this context. The main aim of proteomics is the study and the analysis of the proteins present inside a biological sample or tissue, to characterize it in detail. It is well known that the proteins have, at every levels, a key role inside the cell, and they are responsible both for the physiological and for the pathological state of the cell. The main technological instrument used in high throughtput proteomics to analyze biological tissue from a molecular point of view is the mass spectrometer. The mass spectrometer, often coupled with an upstream chromatography column, allows us to have, starting from a biological sample, a three-dimensional signal similar to a map. The two dimensions of the plane are the mass to charge ratio (often labeled as m/z) and the retention time spent by the molecules to elute from the chromatography column. The third dimension, that is the z-axis, is the intensity of the signal. The localization of the peptides (the proteins, before being analyzed, are always digested in smaller parts, the peptides, using an enzyme such as the Trypsin) happens mainly due to the information related with the m/z and retention time axes. The third axis, instead, has the very important information about both the identification and the quantification of the peptides present in the sample. One of the most important objective of proteomics is the precise quantification of the proteins within a biological sample: this is indeed the quantitative proteomics. There are two kind of quantitative proteomics: absolute and relative quantitative proteomics. In the first case the aim is to quantify the amount of proteins in a sample without any reference, but in absolute terms: in this case it is not present any kind of comparison. In the relative quantitative proteomics, which is widely used for operative and functional reasons, the quantification is performed comparing two or

more different samples: in this case the aim is to determine if there is any difference between the compared samples, or, in other terms, if a protein is over-expressed or sub-expressed. It is very important to be able to perform such analysis and quantification in order to establish which are the proteins, or the networks of proteins, directly related with a determined pathological state. As said before, it is possible to perform quantification with more than two samples: for example, in the datasets shown in this work - kindly provided by the Biomolecular Mass Spectrometry and Proteomics group headed by professor Heck -, there are three samples analyzed in the same signal. To get the signals of the same peptide not overlapped in the exactly same area of the three-dimensional map, it is necessary to label differently the peptides coming from the different samples. In particular, there are several kind of labeling, and in this work the datasets have been realized using two kind of different labeling: the dimethyl and the SILAC labeling. This peptide marking process varies the mass of the differently labeled peptides, without varying their chemical properties. In this way it is possible to visualize in the final signal the couple (or the triplet) of the peptides, composed by the distributions of the same peptide but from different samples. The signals are relatively shifted because of the difference in mass due to the labeling. Comparing the two signals, one lighter and one heavier, it is possible to get information about the relative quantification. Two softwares, in particular, perform this kind of analysis, composed by the first step, which is the peptide identification, and then by the quantification of the peptides. The first software, freeware and widely used, is named MaxQuant, and it has been realized in the Max Planck Institute of Berlin. The second program, commercial and under license, is named Proteome Discoverer and it is sold by the ThermoScientific company. Both these software have been used in this work on the considered datasets, and it is present a comparison between their performances, analyzing the differences. In particular, Proteome Discoverer seems to be much more effective in the identification process, while in the quantification part the program obtain comparable results.

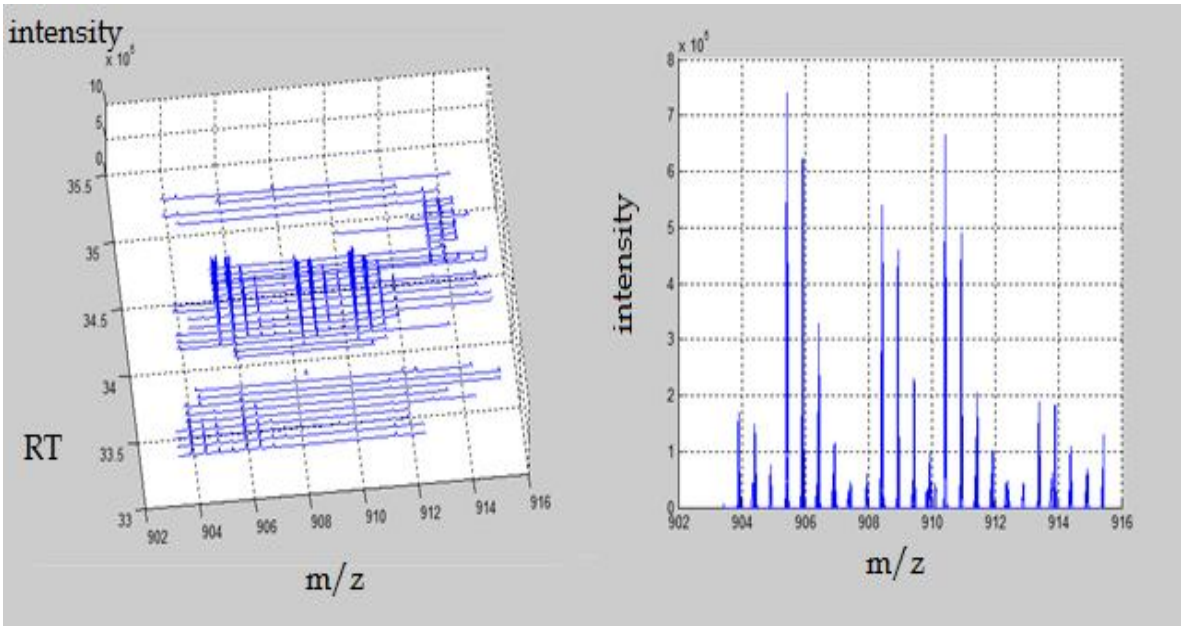


Fig.1 Example of triplet in the three-dimensional map. To the right the same peptide visualized bi-dimensionally

Starting from the identification performed by these programs, our algorithm's aim is to increase significantly the precision and the accuracy of the results obtained using MaxQuant or Proteome Discoverer (mostly MaxQuant, widely used). In particular, we have focused our efforts on two typical problems related to the quantification process in mass spectrometry, which invalidate and worsen the results. The first problem is about the overlap between features of the same peptide, labeled differently: when the shift along the m/z axis isn't long enough, the signals overlap, altering the final ratio of quantification. The second problem, instead, is related to the co-elution of different peptides in approximately the same position in the three-dimensional map: the overlap between different signals generates obviously artifacts which invalidate the quantification. In particular, this problem is not directly tackled by MaxQuant, and this lead to some cases of poor quantification.

The ideas used to solve such problems are shown in the following. Once the area of the peptide is identified in the m/z and retention time axis, a scan selection is firstly performed. Scan by scan, it is computed the Pearson correlation with the theoretical distribution of the peptide and, if the coefficient is higher than a fixed threshold, the scan is kept for the final quantification, otherwise it is discarded. In

this way, a first filtering step is performed to discard the corrupted or noisy information. The second idea concerns the division of the elution area in several sections, both along the retention time and along the m/z axis, in order to get different ratios for the same peptide. Once there are several ratios, these are compared, and only those ratios whose difference is lower than a fixed threshold are kept for the final quantification. If there are two groups of ratios (whose difference is lower than the threshold), for the final quantitation is kept the longest group, or those with a smaller difference. In this way, it is used only that part of the area whose information about the relative quantification is coherent between the sections.

Finally, another original idea introduced in this work, is about the classification of the peptides quantified. In particular, each quantified peptide has a score based upon three different characteristics of the peptide itself:

1. The identification score of the peptide (provided directly by MaxQuant or Proteome Discoverer);
2. The result of the Pearson correlation between the peaks along the retention time axis;
3. The number of sections used for the final computation.

In this way, it is possible to have a score for each peptide, in order to rank the reliability of the quantification performed. The three components of the score are weighted according to a linear classifier where the quantification has been considered successfully if the ratio is in the range of the expected value more or less the 50% (which numerically means that the value should be between 0.5 and 1.5, being the expected value equal to 1).

Finally, about the overlap between the elution areas of the same peptide differently labeled, it has been tested a method proposed by a Korean researcher in the 2010, it uses the quadratic equations to solve the overlap issue and get the expected ratio. To evaluate the effectiveness of such method, it has been used a specific dataset with the dimethyl labeling, that shows this kind of trouble.

The results obtained with our algorithm are very interesting. In the case of overlap between areas of the same peptides (dimethyl dataset), our performance is much better than that of MaxQuant. In particular we reduced the standard deviation of

the results of an order of magnitude, keeping almost the same number of quantified peptides. Furthermore, other four datasets have been used, where the complexity of the signal was very high, due to the elevated concentration of peptides. Even in this situation, our algorithm has performed greatly, reducing the standard deviation of the final results (respect MaxQuant performance) in every dataset (expect one, where the values are comparable), and keeping the averaged value close to the expected one. After the scoring, the accuracy of the results further increases, but the number of quantified proteins is significantly reduced.

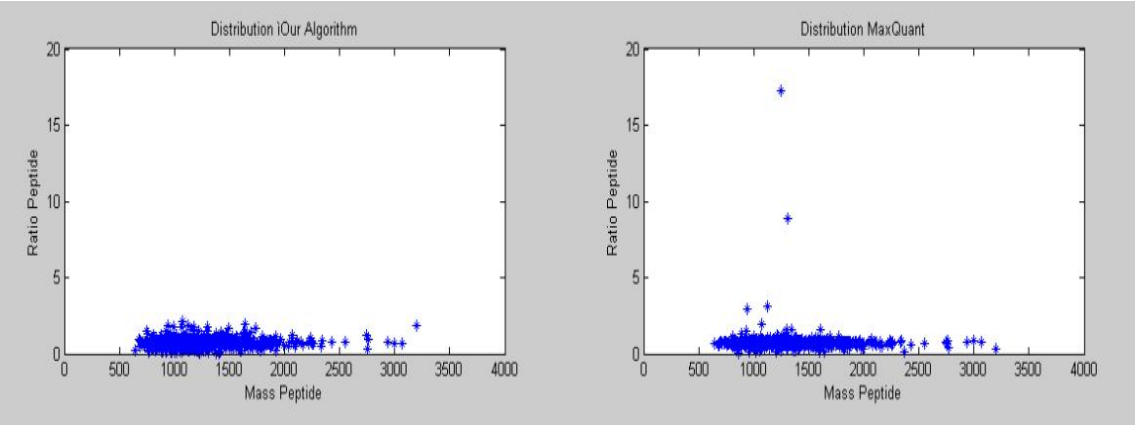


Fig.2 Comparison between our algorithm and MaxQuant results; Dataset SILAC02 1:1:1

There are even reported some specific cases, where it is shown how our algorithm performs much better than MaxQuant in some specific situation. In particular, when an external overlap occurs (specifically with an elution area of another peptide), MaxQuant quantifies providing a completely wrong value. Our algorithm, instead, as shown in the figure, is able to correct such mis-quantifications, approaching the expected value, or avoiding the quantification, due to the lack of information to get the final ratio. In this way, the results are close to the expected value (equal to one in the picture), avoiding outliers far from the expected value (as shown in the picture).

Sommario

La bioinformatica, nell'ambito delle scienze biomediche, sta assumendo un ruolo sempre più importante nel ottenere e nell'interpretare informazioni utili ai fini di comprendere i meccanismi e il funzionamento dei processi cellulari. In particolare, la proteomica ha assunto negli ultimi anni un ruolo sempre più importante in questo contesto. Il compito principale della proteomica è quello di studiare e analizzare la componente proteica presente all'interno di un campione o di un tessuto, per poterlo caratterizzare dettagliatamente. E' noto che le proteine svolgono, a tutti i livelli, un ruolo chiave all'interno della cellula, sia per il suo normale funzionamento fisiologico sia in caso d'insorgenza di stati patologici. Lo strumento principe della proteomica, che ha assunto da anni un ruolo guida nello studio di tessuti biologici dal punto di vista molecolare, è lo spettrometro di massa. Lo spettrometro è uno strumento che, accoppiato con una cromatografia a monte, permette di ottenere, partendo da un campione biologico iniziale, un segnale tridimensionale simile ad una mappa. Le due dimensioni del piano sono la massa sulla carica (tipicamente indicata come m/z) e il tempo di eluizione delle molecole dalla cromatografia a monte dello spettrometro. La terza dimensione, ovvero l'asse verticale z, è l'intensità del segnale. I primi due assi sono utilizzati principalmente per la localizzazione dei peptidi (le proteine, prima di essere analizzate con questa metodologia, vengono quasi sempre digerite in parti più piccole, denominate peptidi). Il terzo asse invece contiene informazioni preziose sia per quanto riguarda l'identificazione, sia per quanto riguarda la quantificazione dei peptidi nel campione analizzato. Uno dei principali obiettivi della proteomica, difatti, consiste nel definire precisamente la quantità di componente proteica presente all'interno di un campione: in questi casi si parla di proteomica quantitativa. Ci sono fondamentalmente due tipologie di proteomica quantitativa: assoluta e relativa. Nel primo caso si tenta di stabilire qual è la quantità di proteine presenti in un campione in termini assoluti, senza effettuare nessuna sorta di comparazione. Nel secondo caso invece, ampiamente più utilizzato sia per

questioni di fattibilità operativa sia di tipologia di informazioni ricavate, la quantificazione viene eseguita comparando due tessuti o campioni diversi: in questo caso si tenta di stabilire se nei due campioni esistono delle differenze quantitative, ovvero se una proteina è più sovra-espressa o sotto-espressa in uno dei due campioni. Ovviamente, da un punto di vista clinico, è fondamentale essere in grado di eseguire questa tipologia di quantificazione per poter stabilire quali sono le proteine legate a determinati stati patologici: difatti, frequentemente la proteomica quantitativa relativa viene eseguita comparando un campione di tessuto fisiologico con un campione dello stesso tessuto patologico. In questo modo è possibile individuare le proteine, o la rete di proteine, responsabili dell'insorgenza della patologia. Inoltre, oltre che a quantificazioni binarie tra stati fisiologici e patologici, questo tipo di quantificazione si presta anche ad altre tipologie di analisi, e può essere effettuata anche con più di due campioni: i dataset presentati in questo lavoro ad esempio, - gentilmente forniti dall'università di Utrecht e in particolare dal gruppo Biomolecular Mass Spectrometry and Proteomics guidato dal professor Heck, dove parte di questa tesi è stata svolta- presentano una comparazione tra tre campioni diversi. Ma come avviene esattamente la comparazione tra campioni diversi? Per poter ottenere dei segnali non sovrapposti nella mappa fornita dallo spettrometro, è necessario etichettare (dal termine inglese 'labeling') i peptidi provenienti dai vari campioni in maniera diversa. In particolare, esistono diverse tipologie di labeling, e in questo lavoro si sono affrontati dataset realizzati con due diverse tecniche: il labeling dimetile e quello SILAC (basato su isotopi). Questo processo di marcatura dei peptidi permette di variare la loro massa, senza alterare le proprietà chimiche della molecola. In questo modo sarà possibile visualizzare nel segnale finale una coppia (o tripletta) di segnali appartenenti al medesimo peptide, non sovrapposte in virtù dello spostamento lungo l'asse della massa su carica, e provenienti dai due campioni diversi. Comparando i due segnali è possibile ottenere le informazioni relative alla quantificazione.

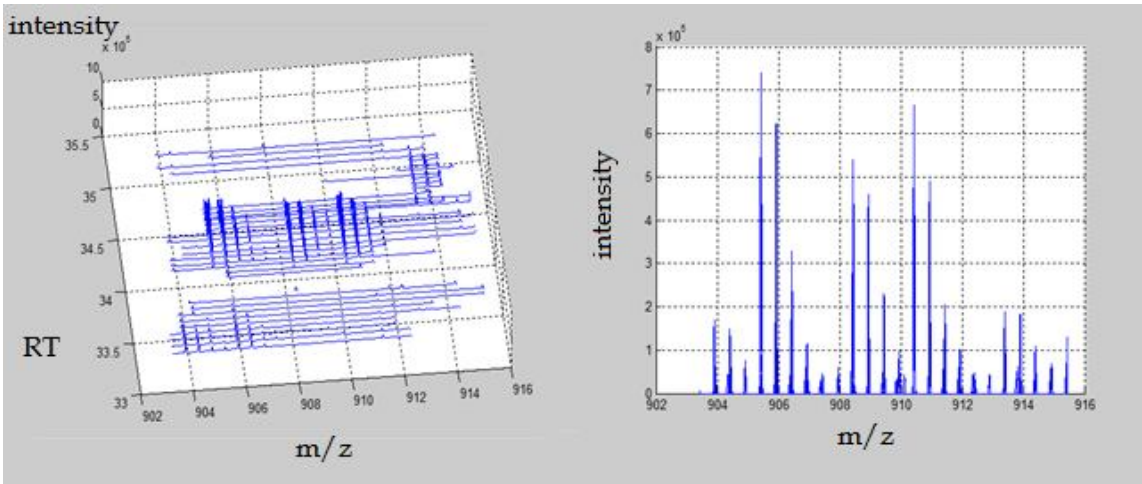


Fig.1 Esempio di tripletta nella mappa tridimensionale. A destra lo stesso peptide visualizzato bidimensionalmente

Due software in particolare si occupano di questo tipo di analisi, ovvero identificazione e quantificazione relativa di campioni analizzati tramite spettrometri di massa. Il primo, ampiamente diffuso in virtù della sua disponibilità freeware, è MaxQuant, ed è stato realizzato al Max Planck Institute di Berlino. Il secondo software, commerciale e distribuito dalla ThermoScientific, si chiama Proteome Discoverer. Entrambi questi programmi sono stati eseguiti sui dataset utilizzati in questo elaborato, e un'analisi delle relative performance indica una prestazione migliore nettamente di Proteome Discoverer in fase di identificazione, mentre i risultati sono comparabili in termini di quantificazione, anche se ancora una volta Proteome Discoverer si mostra essere leggermente più accurato.

A partire dall'identificazione eseguita da questi software, il nostro algoritmo ha l'obiettivo di aumentare la precisione e l'accuratezza dei risultati ottenuti utilizzando MaxQuant o Proteome Discoverer. In particolare, ci si è focalizzati su due problemi tipici che peggiorano sensibilmente il risultato della quantificazione (in particolare di MaxQuant, la cui documentazione è disponibile, ed è nota la strategia utilizzata per quantificare). Il primo problema riguarda l'overlap tra peptidi uguali ma marcati in maniera diversa: quando lo spostamento lungo l'asse della massa non è abbastanza consistente i due segnali si sovrappongono, alterando così il rapporto di quantificazione cercato. Il secondo problema invece riguarda la co-eluizione di peptidi diversi nello stesso punto della mappa proveniente dallo spettrometro: la sovrapposizione dei segnali genera ovviamente artefatti che

inficiano e falsificano la quantificazione. In particolare quest'ultimo problema non è direttamente affrontato da MaxQuant, e ciò comporta alcuni palesi casi di mal quantificazione.

Le idee utilizzate per risolvere questi problemi sono esposte nel seguito. Identificata l'area di eluizione del peptide in esame, inizialmente si esegue una selezione di ogni scan componente il segnale stesso, utilizzando il coefficiente di Pearson per ottenere una correlazione tra il segnale sperimentale e la distribuzione teorica nota. Se la correlazione non supera una certa soglia, lo scan non viene considerato per la quantificazione. In questo modo si attua una prima fase di filtraggio dell'informazione utile per la quantificazione. La seconda idea riguarda la divisione dell'area di eluizione in diverse aree, di modo da ottenere per lo stesso peptide diversi rapporti. Una volta che sono stati ottenuti diversi rapporti, sia lungo l'asse del tempo di eluizione, sia lungo l'asse del rapporto massa su carica, questi vengono confrontati tra di loro. Se l'informazione è coerente (ovvero la differenza tra i rapporti è minore di una data soglia) allora viene eseguita la quantificazione; se invece ci sono gruppi di rapporti molto diversi (in virtù di una sovrapposizione o di una qualsiasi tipologia di rumore) viene utilizzato per la quantificazione quel sottogruppo di rapporti, se presente, la cui differenza è minore della soglia. In questo modo si isola la parte di area il cui rapporto (ovvero la sua quantificazione) è diverso rispetto al resto del segnale.

Infine, un'ulteriore e originale idea implementata nell'algoritmo presentato, riguarda la classificazione della quantificazione eseguita. In particolare, per ogni peptide quantificato, viene assegnato un punteggio sulla base di tre caratteristiche del peptide stesso:

1. Il suo punteggio di identificazione (direttamente fornito da MaxQuant o Proteome Discoverer)
2. Il risultato della correlazione di Pearson ottenuta tra i picchi lungo l'asse del tempo di eluizione (si ricordi che la selezione degli scan viene effettuata utilizzando il coefficiente di Pearson lungo l'asse massa/carica)
3. Il numero di sotto-aree utilizzate per il calcolo del rapporto finale

In questo modo, è possibile per ogni peptide avere un punteggio che stabilisca qual è l'affidabilità (ovvero la probabilità) che il risultato di quantificazione fornito

sia corretto. Il passo successivo è ovviamente quello di pesare appropriatamente i diversi contributi, e ottenere un punteggio in grado di stabilire se il peptide è stato ben quantificato oppure no. Attraverso un classificatore lineare è stata realizzata una classificazione supervisionata, nella quale la quantificazione viene considerata avvenuta con successo se il valore ottenuto è compreso in un dato intervallo, pari al valore del rapporto atteso più/meno il 50% (che si traduce numericamente nell'intervallo 0.5 < rapporto < 1.5, essendo il rapporto atteso uguale a 1 per il dataset usato come training set). In questo modo si ottiene una quantificazione dai risultati estremamente accurati, con l'ovvio compromesso di escludere alcune proteine dalla quantificazione finale. Per quanto riguarda infine la sovrapposizione tra le aree dello stesso peptide marcate differentemente, è stato utilizzato un metodo presentato da un ricercatore coreano nel 2010, che sfrutta la risoluzione di equazioni quadratiche per ottenere il rapporto corretto. Per valutare l'efficacia di questo metodo è stato utilizzato un dataset in cui il labeling utilizzato (dimetile) presentava questa tipologia di problema.

I risultati ottenuti sono molto interessanti. Nel caso della sovrapposizione tra aree dello stesso peptide differentemente marcate, la nostra quantificazione si è rivelata essere di gran lunga migliore rispetto quella di MaxQuant, riducendo la standard deviation dei risultati ottenuti (circa 1500 peptidi che presentavano la sovrapposizione) di un ordine di grandezza, e mantenendo comunque un buon numero di peptidi quantificati.

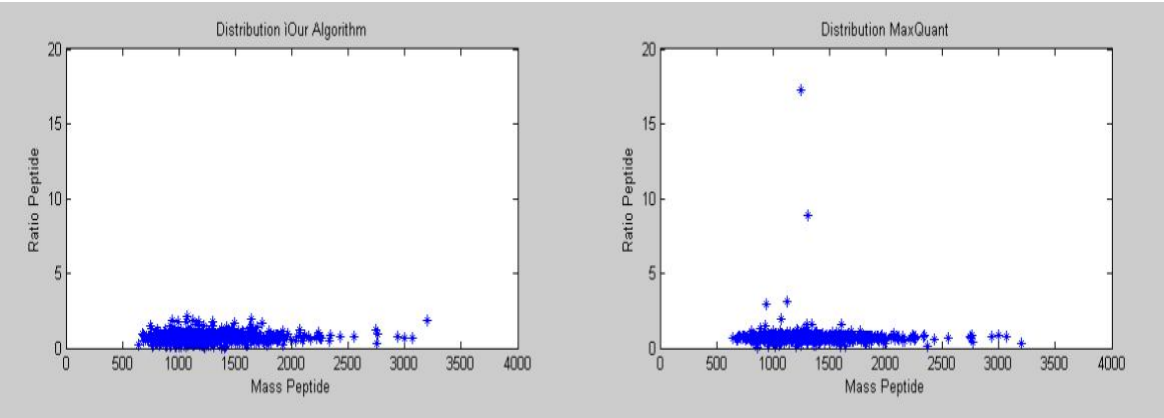


Fig.2 Paragone risultati nostro algoritmo vs. MaxQuant; Dataset SILAC02 1:1:1

Sono stati inoltre utilizzati altri quattro dataset, nei quali era presente una rilevante complessità dovuta elevato numero di peptidi presenti nel segnale. Anche in questo caso, i risultati della quantificazione ottenuti con il nostro algoritmo hanno ridotto significativamente la standard deviation dei risultati in tutti i dataset (eccetto uno, dove i risultati sono comparabili), ottenendo così dei risultati migliori rispetto quelli ottenuti da MaxQuant. Nella tesi, inoltre, sono riportati dei casi specifici nei quali MaxQuant, in presenza di una sovrapposizione esterna, quantifica in maniera completamente erronea, introducendo nei risultati dei valori chiaramente 'outlier'. Il nostro algoritmo invece, come mostrato nella figura, è in grado di correggere queste quantificazioni, o riportando il valore corretto, o non quantificando affatto, per via della sovrapposizione che non permette il calcolo del rapporto in maniera esatta. In questo modo, i risultati sono centrati attorno al valore atteso (pari a uno nella figura) e senza alcun valore palesemente sbagliato.

Introduction

Nowadays Proteomics has reached a pivot role in Biology and Medicine: proteins are actually the structure, the engine and the operative components of the cells.. In clinical biology, quantitative proteomics is performed to properly understand which proteins are involved in different physiological or pathological states. In this context, where the research plays a primary role in the medical framework, it is often performed the relative quantification between two opposite states, such as health/diseased or sample/control. To perform this kind of relative quantification process, usually, the proteins of the first sample are labeled with a light label, while the proteins from the second sample are labeled with an heavier label, in order to have the same peptide differently marked. In this way, it is possible to compare the signals coming from the two samples, which are separated due to the mass shift, using for example a chromatography coupled with a mass spectrometer (LC-MS). In the mass spectrometry, the mass of the peptide has an important role: indeed, in LC-MS, peptides are represented in a three-dimensional space, where the axes are the mass to charge ratio and the elution time (usually named retention time), which is the time spent by the peptide to elute out from the chromatography column. The last axis is the intensity of the signal (whose information is that one used to perform the quantification). Nowadays mass spectrometers provides data with a very high resolution, and allow the identification and the relative quantification of a very high number of proteins in a sample. For these reasons, the actual state of art of the research in proteomics can't disregard the contribute provided by the mass spectrometry. The data obtained with the mass spectrometers are actually very rich, and the relevant biological information may be sometime easily inferred, sometime it requires elaborated processing to give precious biological notions (such as, once again, the quantification of the amount of proteins in a sample). One of the bioinformatics' roles in proteomics is essentially to extrapolate, as much as possible, the desired information from the data, in a trustable and reliable way.

This work is about a post processing algorithm, whose aim is to provide information about the relative quantification of labeled peptides of a sample, coming from a mass spectrometry experiment. Generally, a software used for mass spectrometry-based experiment can be basically divided in two steps: the identification of the peptides, and their quantification. Our algorithm quantify the peptides with an innovative approach, starting from the identification performed by two different programs, which are the actual state of the art in the identification and quantification processes (MaxQuant – freeware and widely used, and Proteome Discoverer - under license). The main idea at the basis of the algorithm is to combine two different strategies to achieve better results in the quantification, compared with the results obtained by the software used for the identification process. The first one is to exploit the information related to the identification process: each scan is compared with the theoretical distribution of the peptide through the Pearson's coefficient and, if the result is over a predetermined threshold, the experimental distribution is kept for the quantitation; otherwise the scan is discarded. In this way, scans selected for quantification are those, in principle, properly shaped and well identified, whose information about the quantification of the peptide is reliable. The second idea implemented in this algorithm is to divide the bi-dimensional area of the peptide (mass to charge and retention time dimensions) in several sections, after have performed the scan selection. In this way it is possible to compute different quantitation ratios for the same peptide and, matching them, it is possible to leave out those -if present- which differ from the others. Thereby it is possible to avoid those sections which are affected by overlap with the elution area of other peptides, or those areas which are corrupted by noise. This idea has been mainly thought to solve the problem of the overlap between signals from different peptides: this problem is relevant due to the high concentration of proteins in a biological sample. The big number of proteins (and, obviously, of peptides in the final signal) in the sample increases the complexity of the data itself, and misleads the quantification process, because of such overlaps between signals of different peptides. The second idea aims to provide a solution to this relevant problem. Another issue related to the overlap between peptides, and that has been tackled in this work, is

about the overlap between features of the same peptide. Some kind of labeling may lead to an overlap between the Light area and the Heavy area, introducing a bias in the quantification. This may happen because the label introduces a mass shift which is pretty short, and the first peak of the heavy peptide overlaps with the other peaks of the light peptide (as we will see, the isotopic distribution of a peptide is composed by several peaks). In this work two versions of the same algorithm are shown, these are slightly different because optimized for the kind of labeling used in each datasets. In the first dataset it has been implemented a method *ad hoc* to solve the overlap problem between the Heavy and the Light distribution, while in the other datasets, where the problem of the co-eluting peptides is very present, it has been increase the number of sections in which the elution area has been divided, and the quantification strategy has been slightly modified to better solve that specific problem.

Finally, a scoring process has been introduced, in order to rank the quantification process of each peptide. In particular, the score is given using three characteristics of the peptide and of the quantitation process.

In conclusion, the aim of this work is the implementation of a workflow able to improve the quantitation results obtained with the best state of the art algorithm, such as MaxQuant. In particular, we aim to improve the quantification of those peptides whose complexity, due to different kinds of overlap with other distributions, would lead to a significantly poor quantification.

Every dataset used in this work have been kindly provided by the Bio Molecular Mass Spectrometry and Proteomics laboratory of Utrecht, headed by Professor Heck, where part of this thesis has been developed.

Overview

It was 2008, when Nature Biotechnology published an editorial named: ‘Prepare for the deluge’ [1], where the author correctly forecast the burst of available data in the field of biology. Actually, that paper was about genomics data, but it’s easily extendable to the proteomics field, and to the all newborns ‘–omics’ fields that are going to live a new renaissance in biology and biotechnology. In order to properly understand and integrate this huge amount of data in an automatic and comprehensive way, a new discipline is moving its first steps: bioinformatics. The aim of bioinformatics, as already stated, is the analysis of the data provided by the new technologies (such as microarrays or mass spectrometers as well), and the extrapolation from such data of new information useful to understand the mechanism of the living cell. This thesis could find its location in such field: elaborate and analyze raw data provided by a mass spectrometer, from a sample realized on purpose, in order to be able to extract a very specific kind of information; to be more specific, the information is about the relative quantitation of proteins, at a peptide level. To understand the context in which this thesis is developed it is important first to glance at the proteomics field in general, and then deepen in detail the quantitative proteomics and the instrumentation used for this task. Let’s then start this chapter with an overview about proteomics in general.

Proteomics

Proteomics may be defined, at first, as the study of a subset of proteins present in a specific part of the organism, and how these proteins change during time and varying conditions. We can summarize the huge field of proteomics in four main cornerstones, which enclose all the different subfield of such discipline:

1. *Protein Identification*: it is the determination of which proteins are present in a sample, separating and identifying uniquely each protein. To do so, it is

important to know either the sequence, or so many physical characteristics that it is statistical unlikely that the protein could be another one;

2. *Protein Characterization*: it's the determination of the biochemical and biophysical characterization of the protein, although the protein itself may not have been identified yet;
3. *Protein Quantification*: it is the determination of the amount of proteins present in the sample. It may be two different kinds of quantification: absolute and relative. The first one is much more difficult to be reached, and it may be reduced to the relative quantitation between the sample and some internal standard. The second one, as we will see in the following of this chapter, is somehow easier to be achieved (but far from being trivial). In this work, we are going to focus on the relative quantification between three different samples;
4. *Sample Comparison*: it's somehow the unification of all the other three points, and it determines the similarities and the difference in the protein composition of two different samples. Some aspects may be the relative occurrence of the proteins, the relative abundance or the presence of some differential modification.

Quantitative Proteomics and the Experimental Workflow

In the proteomics field, the quantitative task is very important in order to get the expression of a protein in two different samples, related by a Boolean state such as healthy/diseased, or young/aged: in this way, for example, it is possible to understand which are the proteins related to the studied disease, because over-expressed or sub-expressed, and therefore understand the network of interactions at the basis of the disease. As it is possible to see in the figure 3, the general workflow is divided in about five parts: protein *isolation* from the sample, protein *separation*, protein *digestion*, peptide *fractionation*, mass spectra analysis and finally *data analysis*. Once the proteins have been separated through gel techniques, it is possible to perform the protein digestion, in order to get the peptides from the proteins. The enzymes that performs the digestion are called

proteases and these are chosen in order to cleave the peptides in a very predictable and consistent way. It is important that the obtained peptides are not very long or too short, because often the mass spectrometers have a limited mass range, beyond which is useless to have any sample.

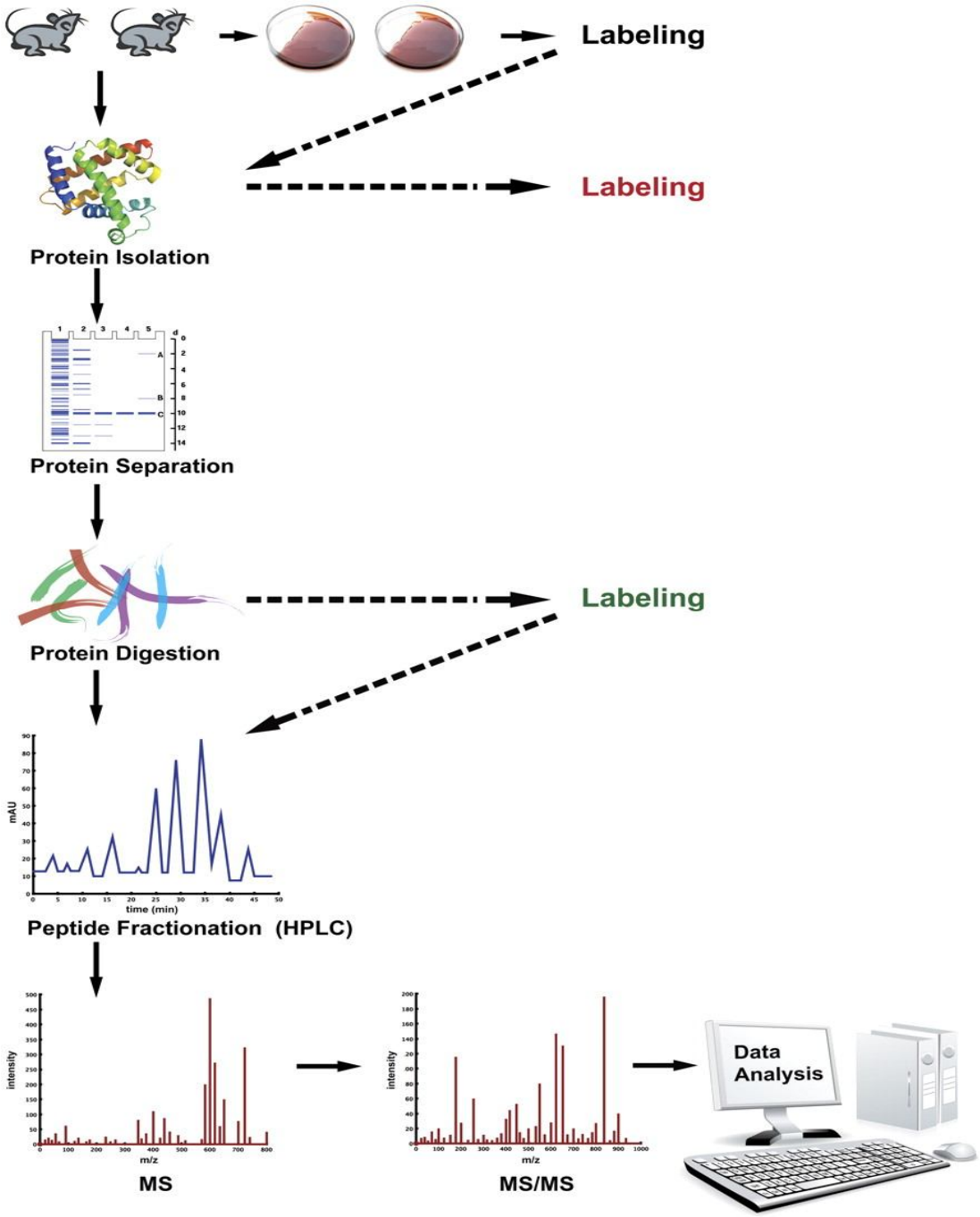


Fig.3: Quantitative Proteomics workflow

Furthermore, it's important not to cleave the peptides in short parts, because the

higher is the number of the peptides, the higher is the complexity of the signal and the worse is the identification and the quantification. It is therefore important to find out the right trade-off. This compromise is reached by the protease known as *trypsin*, which offers some advantages, such as:

- High specificity, which means very low rate of miss cleavage, and virtually no cleaves at unexpected position;
- The arginine and lysine appears in the proteins approximately every 11 residues: the peptides produced are of suitable length for mass spectrometry analysis;
- The trypsin is easily obtained and purified.

Once the proteins have been cleaved in peptides, it is performed the Liquid Chromatography step, where the eluate run in a column with predetermined chemical features, which retains differently the running peptides. Therefore the chromatography is a way to get a further division in the sample, based on the retention time: as the eluate comes out from the column, it is analyzed by the mass spectrometer. Normally, an experimental run in chromatography, coupled with a mass spectrometer, is performed with a order of magnitude of hours. In the mass spectrometer, the sample is analyzed at different levels: in the first one it is possible to get the signal of the whole peptide, in the further levels the peptide is fragmented and analyzed in order to get spectra useful for the identification purpose. In the next paragraph will be given more details about the mass spectrometer and the dataset provided. The next question about quantitative proteomics comes directly from the figure 3 that shows the workflow: how exactly works the quantitation process and what is the labeling?

There may be two different kind of strategy: the first one is the label-free quantification, while the second one is the label-based quantitation. The label free method is based on the comparison of different subsamples, coming from the samples that are going to be analyzed. Once the sample has been digested and has run in the chromatography column, it is possible to get the mass spectra, with the intensities for each peptide. At this point it is performed the comparison of the signals from the mass spectrometer: it is necessary to find out the correspondences of the spectra matching the signals obtained. This is the most

difficult part of the experiment, from a computational point of view.

On the other side, there is the label-based quantitation, which is the one used in the realization of the dataset used in this work. The label based method can be divided in two parts: the first one is performed by the MS based quantitation, the second one required the second level of analysis that is the MS/MS based quantitation. The main idea is that the peptide molecules are labeled differently for each sample: with a Light molecule, a Medium and an Heavy one; since we know exactly the mass of the label used, and where the label is performed in the sequence, we are able to find for each peptide the whole triplet (Light – Medium – Heavy). Comparing the intensities of the peaks of each feature, we should be able to quantify the differently labeled peptides. In the figure 4, it is shown an example of triplet, composed by the isotopic distribution of the Light labeled peptide, the Medium and the Heavy one.

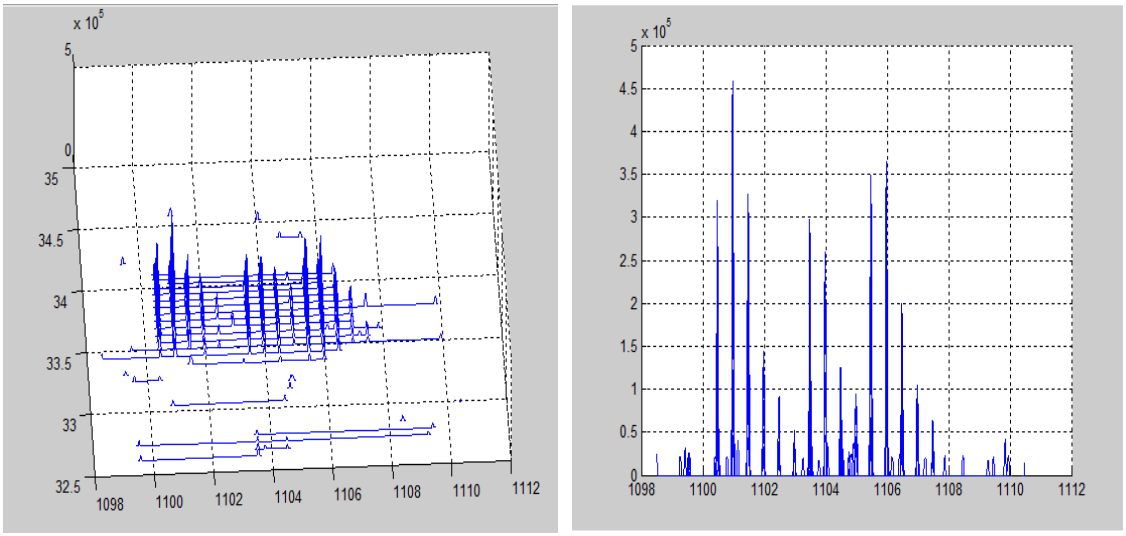


Fig.4: MS Signal: a well defined Triplet

Finally, the last type of quantitation, developed in the proteomics field, is the absolute quantitation. Absolute quantitation is often expressed as the molar concentration of a protein in a sample. One example where the absolute concentration of a given peptide can be of interest is when looking for biomarkers, where the absolute concentration of a peptide biomarker will provide useful information about the suitability of different assays to detect this peptide in a

subsequent diagnostic procedure. Often, this kind of quantitation is a relative quantitation, where the corresponding peptide is an internal standard used as a meter for the quantitation process (so, in this sense is not totally correct to speak about absolute quantitation).

Mass Spectrometry: instrumentation and data provided

The Mass Spectrometer (MS, fig.5) is the main instrument used in proteomics, and particularly in quantitative proteomics. Before understanding how it works, we briefly see how it evolved from the first steps of its life till our days. The first documented application of mass spectrometry to a proteomics experiment dates back to 1958 [4], thanks to the efforts of the pioneer Carl-Ove Andersson, who worked with the fragments ions of the methyl esters. At the beginning of the century, precisely in the 1918 and 1919, Arthur Jeffrey Dempster and F.W. Aston worked on the implementation of some modern techniques used in mass spectrometry. Many decades later, in 1989, Hans Dehmelt and Wolfgang Paul were awarded of the Nobel Prize in Physics for the development of the ion trap technique (a work carried out in the 1950s and 1960s). Then, in the 2002, John Bennett Fenn and Koichi Tanaka won the Nobel Prize in Chemistry for the development of the electro-spray ionization (ESI) and the development of the soft laser desorption (SLD) and, obviously, their applications in proteomics. Finally, the *Orbitrap*, a type of Mass Spectrometer with the highest resolution, has been invented by Alexander Makarov, who received for his efforts the American Society for Mass Spectrometry Distinguished Contribution in Mass Spectrometry award in the 2008 [5]. The history of the mass spectrometer is quite short, being this instrument recent: but how does it work?

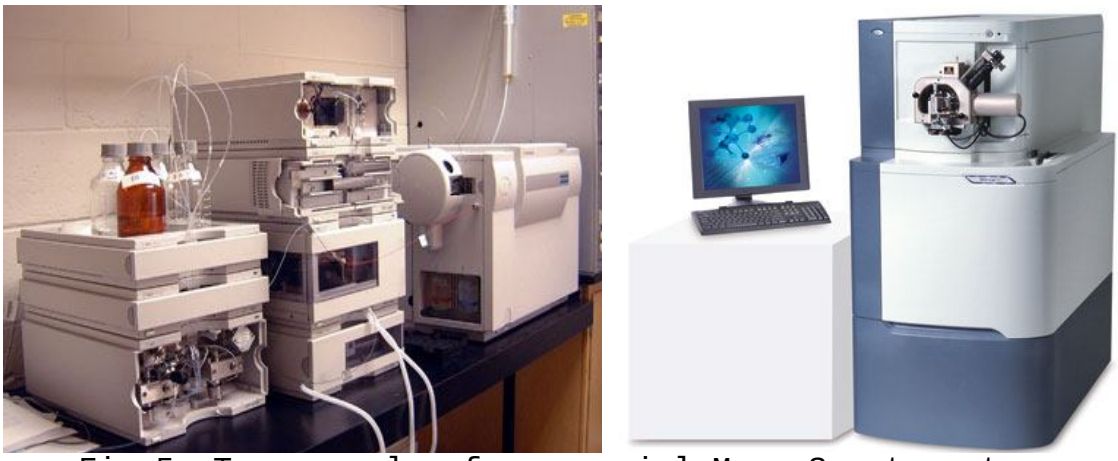


Fig.5: Two example of commercial Mass Spectrometers

Every Mass Spectrometer can be divided ideally in three parts: the ionization source, the mass analyzer and the detector (fig.6). The first component, the ionization source, is required because to handle the components of the sample (peptides, e.g.), these have to be ionized: the mass spectrometer uses electric or electromagnetic forces to move and to measure the components of the sample. Usually, the ionization is achieved by adding protons to the molecules, and there are several ways to obtain this addition. We are going to see a couple of them (the most used). Once the samples are charged, they are transferred to the mass analyzer through the acceleration region, and separated according both to the charge and the mass. After the separation, finally, the charged samples hit the detector, and a mass spectrum may be constructed thanks to a computer connected with the mass spectrometer.

Let's see in detail two kind of ionization source, the MALDI and the ESI. First of all, it is important to stand out the desired features required from a ionization source in proteomics.

- The sample should be ionized in a detectable amount, and the ionized amount should be proportional to the sample components amount.
- There should not be fragmentation of the components when not required, that means that the components shouldn't break into smaller parts which may not be ionized.
- There should be no unwanted adduct ions.

- There should not be ions from other molecules (contaminants).

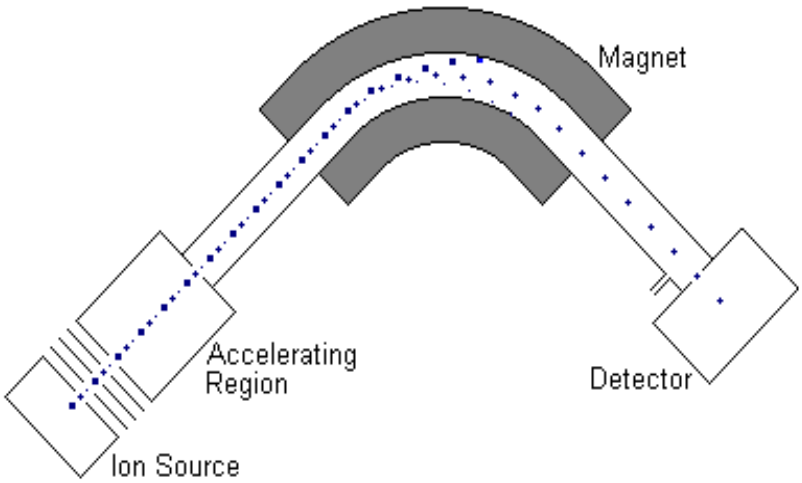


Fig.6: A scheme of a Mass Spectrometer

Furthermore, the sources may be divided in two groups: soft ionization sources and hard ionization sources. The first one causes only limited fragmentation, while the other typically fragment upon ionization. Soft ionization sources are used for peptides and proteins, and if fragmentation is desired afterwards (as in MS/MS) other methods are used to achieve fragmentation. One common ionization source is the MALDI (Matrix assisted Laser Desorption Ionization). The matrix is composed by organic molecules which absorb light in the UV area, and are dissolved in an organic solvent in acidic conditions, then are mixed with the sample. The next step is the evaporation of the solvent, letting the matrix form small crystals, with the sample components incorporated into these crystals (the crystallization process). A pulsing laser (very short pulses of few nanoseconds) is absorbed by the matrix (the wavelength of the laser is the same absorbed by the organic molecules) and therefore by the sample. The matrix has a double role: capture the laser light ionizing the sample, and protect the analyzed molecules from the disruptive energy of the laser. Then, the peptides are able to receive protons from the ionized matrix molecules, and they become ionized in the gas

phase. Most of the ionized peptides carry only one proton. Then, under the influence of an electric field, the ions are transported to the mass analyzer. The environment for this treatment is often under vacuum, or at very low pressure. Otherwise, the ESI is primarily used in the MS/MS analysis. The peptides are brought into the ionization source by a liquid flow. Often, as previously seen, the liquid is the eluate from an HPLC instrument. The liquid is then sprayed into a strong electromagnetic field, and the solvent evaporates, increasing the electric field on the surface of the droplet of the sample (composed by several peptides). When the electric field becomes strong enough, charged peptides desorb from the surfaces of the droplets. Often, the ionized peptides carry more than one proton, and under these conditions are transported to the mass analyzer.

The next step in the ions travel is the mass analyzer. As known, the ions are accelerated by an electric field, and then they enter into a tube. The velocity that the ions have achieved during the acceleration is dependent on the mass and the charge of the ion, and the pass through the drift tube is dependent on the velocity. When the ions hit the detector at the end of the drift tube, the time of flight (TOF) is registered, and the m/z value can be calculated. In this way, it is possible to evaluate indirectly the mass of the peptides analyzed, obtaining the desired spectra, so much useful for our quantification task. The *Orbitrap* mass spectrometer instead, works quite differently. To compute the mass of the peptides it doesn't consider the Time Of Flight of the molecules; but, as shown in the figure 7, the charged peptides spin around the axis of a central electrode, instead of running through the mass analyzer (they are trapped because the electrostatic attraction of the electrode is balanced by the centrifugal force). It is possible to compute the mass of the molecule starting from the frequencies of the oscillations, which are inversely proportional to the square root of the mass to charge ratio. In the commercial version, a linear ion trap can be used as a front end for the *orbitrap*. The accuracy, the resolution and the dynamic range of such instrument, are better than any other instrument nowadays, except for the Q-TOF. The Q-TOF is based on an idea which is similar to the one of the *Orbitrap*. The ions (the charged peptides) are no more trapped in a spinning cycle movement, but they are led through a quadrupole, fig.8, where the electric field forced by the four

electrodes imposed an oscillatory movement to the ions.

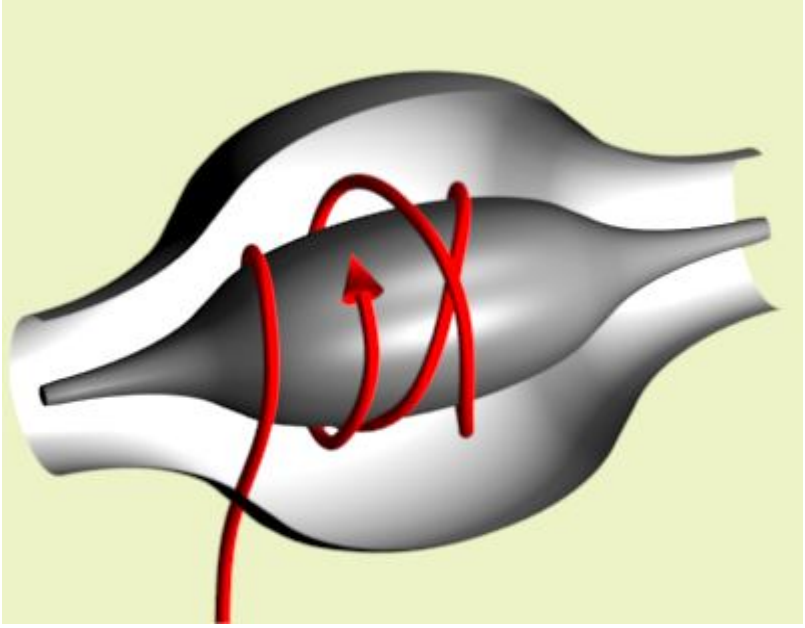


Fig.7: Schematic representation of the *Orbitrap* principle

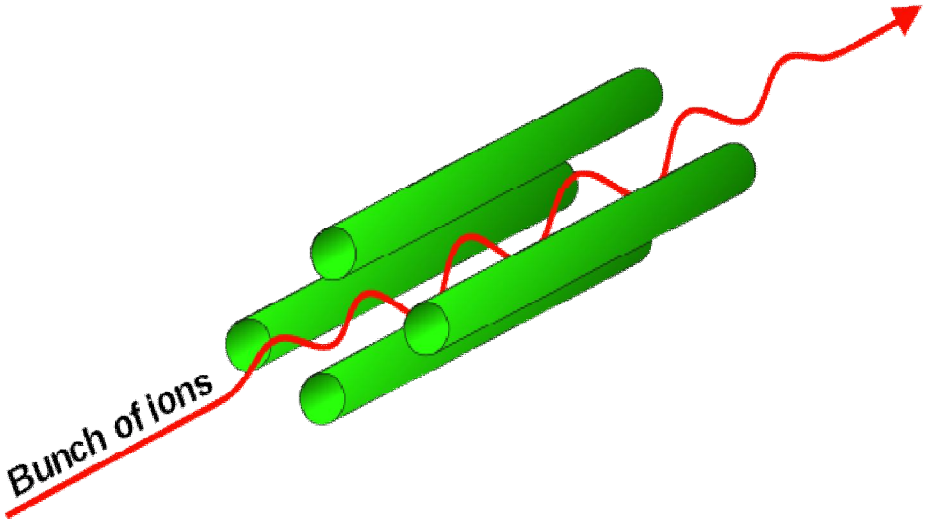


Fig.8: Schematic representation of the *Q-TOF* principle

Therefore, only the stables ions are able to reach the detector, while all the other ions will have an unstable trajectory and they will collide with the rods. This kind of quadrupole is then a very selective and high specific detector. Obviously, tuning

the voltage of the rods, it is possible to scan a wide range of m/z-values.

Raw data obtained from the mass spectrometer may be converted to a more convenient formats, such as .mzXML. This kind of conversion is performed by many workbenches (such as *OpenMS* – see next chapter) or some specific tool (such as *Proteome Wizard*). The mzXML data are easily accessible with any computing language, such as Matlab. The file is composed by three parts:

1. Index, composed by name and offset of each scan;
2. mzXML, with some technical data such as ‘SchemaLocation’ and ‘MsRun’
3. scan, where there are the real data coming from the sample.

In the third part it is possible to get every scan, both from the first and from the second level. Moreover it is possible, for each fragmented ion of the second level, to get the precursor ion, the m/z value, the retention time and the charge. Basically, it is possible to get every information about the fragmented ions, their position and their intensity.

Each peptide has its own elution area, which is spread in the retention time and, because of the isotopic distribution, in the m/z domain. It is possible to see in the figure 9 some examples of this area for some peptides. The dimension of the area is related to the intensity of the signal and then to the amount of the peptide in the sample. The elution profile in the retention time axis should be Gaussian-like, but with a long tail on the second half of the curve. The mathematical functions used is the Boltzmann distribution or the Exponentially Modified Gaussian (EMG). The number of peaks in the m/z axis depends on the intensity of the signal and the level of the noise, which covers the lowest peaks.

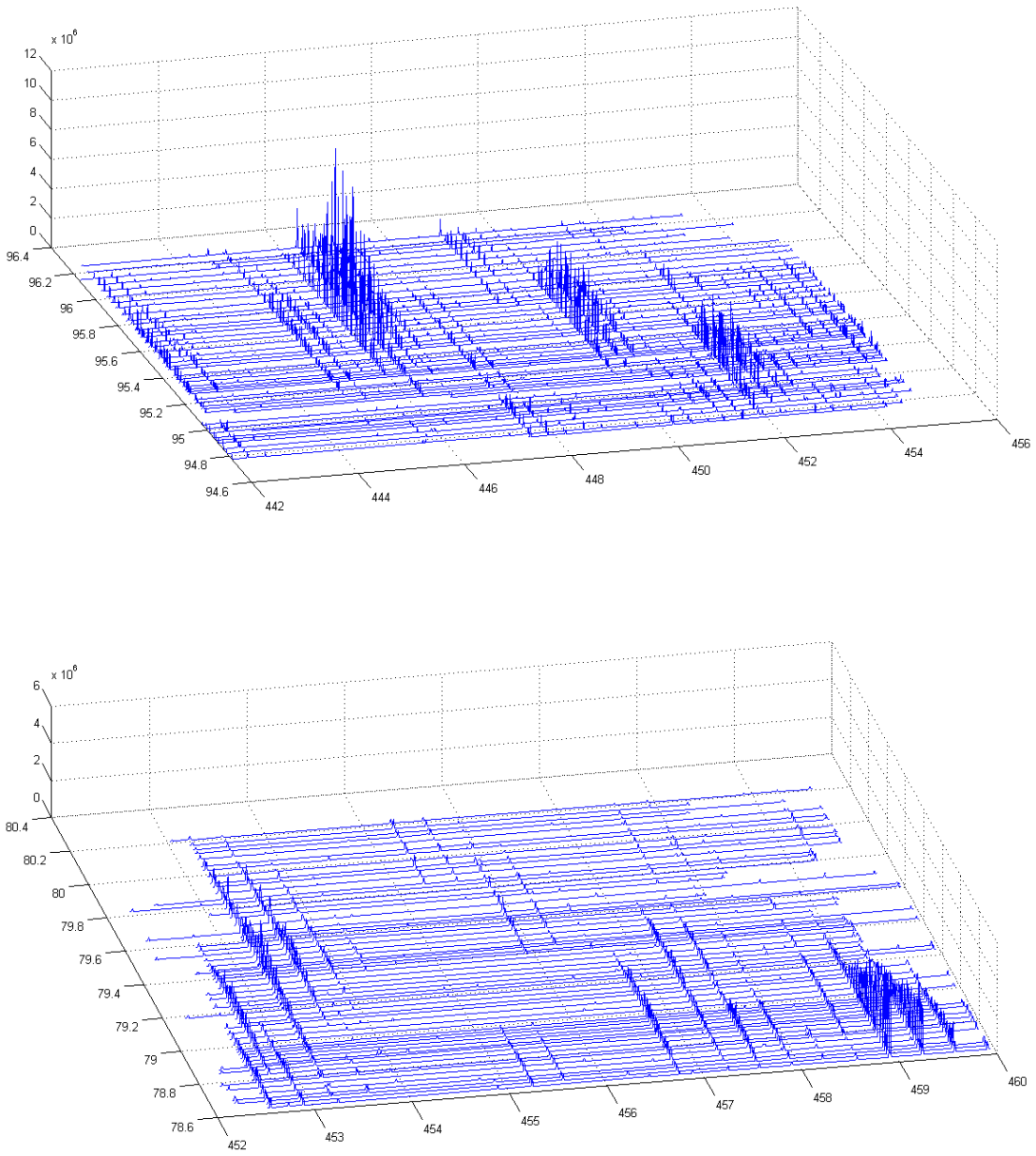


Fig.9: Examples of signals from the Mass Spectrometer with some Peptides.

State of the Art

In this chapter a brief overview will be provided in order to understand how the problem of proteins relative quantification (using Liquid Chromatography coupled with Mass Spectrometry) has been tackled by the bioinformatics community throughout the world, realizing tools and software whose purpose is the automation of the whole process. In the first part of this section some tools, from 2001 to 2007, are briefly described; in the second part of the chapter both the software MaxQuant and Proteome Discoverer have been analyzed in detail; these software are those used in the operative part of this work.

Before starting with the overview of the different methods and tools, it is important to highlight, from an experimental point of view, which are the advantages brought by a Liquid Chromatography - Mass Spectrometry approach in the relative quantification field, and why this approach is worldwide spread in the proteomics field, replacing the previous approach based on gel electrophoresis technique. These are the main strengths of the LC-MS method:

- The sample to be compared (e.g. healthy vs. diseased) are in the same Liquid Chromatography column, therefore there are not problems about different elution time, increasing the reproducibility of the experiment;
- The presence of more dimensions, such as multiple charges for a single peptide, allows a much more precise mass determination, and even better quantification.
- Finally the isotopic labeling doesn't alter the fragmentation process during the first Mass Spectrometry steps.

Starting from these achievements, as previously stated, it would be possible for the scientific community to perform differential analysis in the proteomics expression between cells in opposite state, such as healthy or diseased, in order to understand the mechanism at the basis of the functioning of the cell.

Max Quant

Max Quant (MQ) is a program providing a complete pipeline to analyze data in a quantitative proteomics workflow. In this section, we will refer to the version 1.1.1.36 (fig.1), for the main ideas of the algorithm. Furthermore, this is the version described in detail in the paper of the year 2008 [15]. It is important to remember that this program, though basically the same, has been improved and it has reach the 1.2.2.5 version [16], which is the version used in the operative part of this work. It is possible to divide the workflow in four different parts, which are going to be analyzed in the following:

1. Feature detection and peptide quantification;
2. MS/MS ion search;
3. Identification and validation;
4. Visualization.

In the first part, MQ has to handle the raw data, to get all the necessary information in order to create the isotopic pattern. First of all it is important to locate the peaks: this is done with a local maxima research. As stated in the supplementary notes of the paper [15], *"This straightforward approach of peak detection without any deconvolution, smoothing or de-noising is sufficient for MS data generated by modern high precision mass spectrometer [..]"* In this way it is possible to get the 2D-Gaussian shape for each peak, and, connecting properly in time the centroids of the 2D peaks, we finally have the 3D peak (where the three dimensions obviously are m/z, retention time and intensity of the signal).

Once all peaks in the data have been taken, it is possible to check out which ones stand in an isotope cluster. In order to gather the peaks in a cluster, it is necessary to satisfy three conditions:

1. The difference between the peaks on m/z should be less or equal to a formula containing the bootstrap standard deviations and the maximal shift that the incorporation of a sulphur atom can cause.

2. The intensity profiles should have a sufficient overlap in retention time; to do so it's necessary to compute the cosine correlation (which should be greater than 0.6);
3. The charge of each pattern should be consistent.

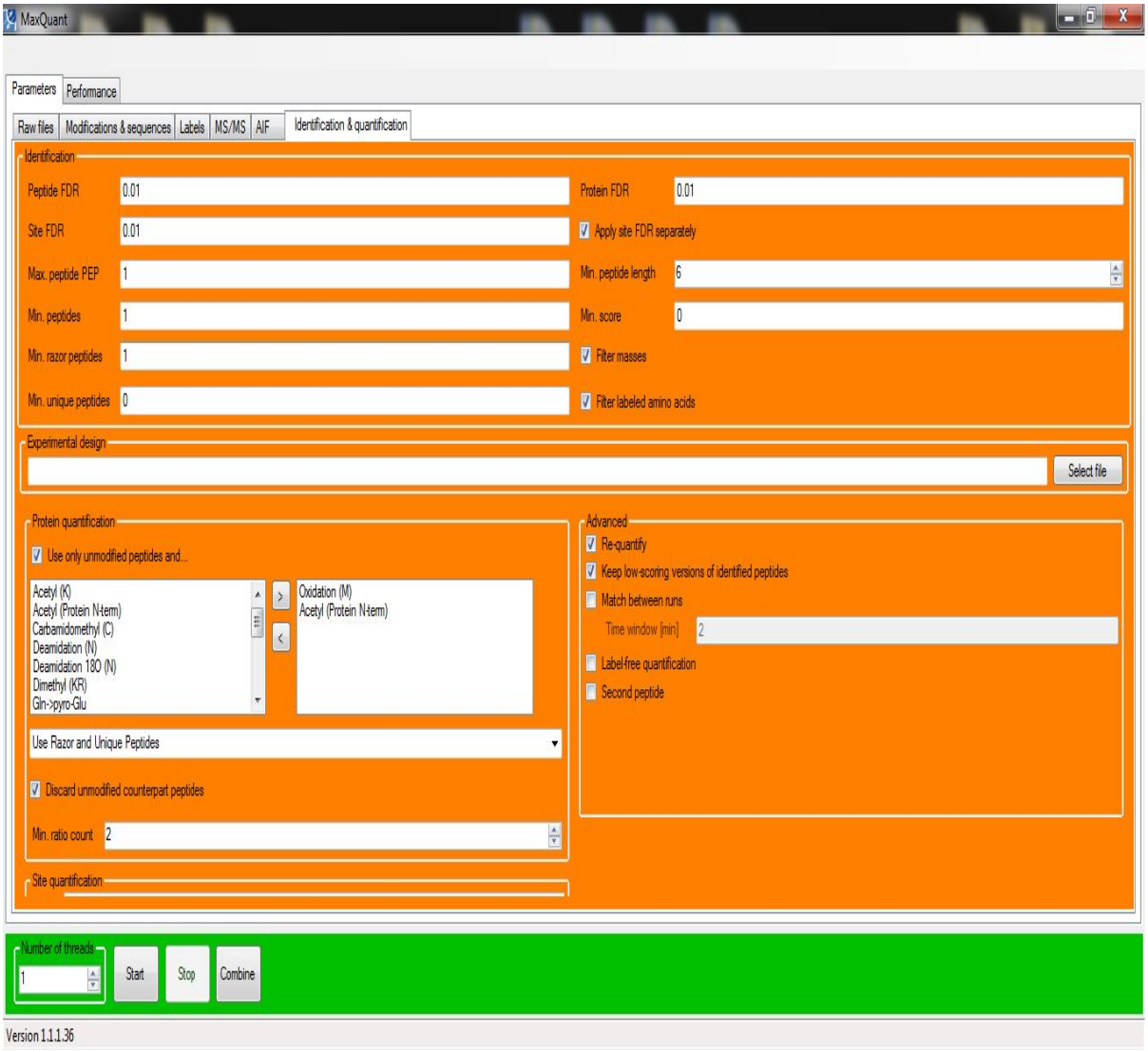


Fig.1: A MQ screenshot about identification and quantification parameters

Now, after all the isotopic patterns have been detected, it is possible to couple them, trying to locate the labeled peptide for each unlabeled one, and vice versa. To create the pair (and, of course, the triplet too) some conditions must be checked: there is a correlation test of the intensity profiles over the retention time, the charge must be the same and, furthermore, the two clusters have to be close enough in mass (clearly the mass shift is label-dependent). Finally the two

measured isotope pattern are convoluted with the theoretical pattern of the missing atoms to obtain the same atomic composition, and therefore the resulting isotope patterns should be exactly identical to less than a factor: that factor is the sought ratio which quantify the abundance of the labeled and the unlabeled peptide. Once the ratios are found, they are normalized so that the median of logarithmized ratios is zero. It's very important for the purposes of this work to point up the strategy adopted by MQ in order to face the overlapping problem. In particular, to avoid erroneous results for overlapping isotope patterns, the ratio calculation is restricted to the first three peaks of an isotope pattern; moreover, if the peptide mass is above 2800 Da, the monoisotopic peak (namely the first one) is excluded for the quantification. For the non-linear recalibration of the mass, it is important to detect the charge pairs, that is those peptides that have been measured in multiple charge states.

The second part of the MQ workflow is where it is used the database engine in order to identify the peptides. In the version of the quoted paper the database engine was Mascot, widely used in the Proteomics context. On the contrary, in the latest versions of MQ, it has been developed a new engine named Andromeda, still used. Before submitting the MS/MS spectra for database search, they are prepared through a filtering phase, and even after the Mascot results are filtered by individual peptide mass errors. This means that those candidates suggested by Andromeda which exceed the mass tolerance (after recalibration) are discarded.

The third part is about identification and validation. After filtering Mascot results and after a linear mass recalibration, two parameters are computed. The first one is named PEP (Posterior Error Probability) and it is the probability of a false hit, given starting from the peptide identification score and the length of the peptide; it is calculated with a Bayesian formulation. The task of the PEP is to determine the second parameter, the FDR, that is the False Discovery Rate. To obtain it, the peptide identifications are sorted from the forward and reverse database by their PEP, and those peptides with 1% of accumulated reverse/forward hits are accepted. Moreover there is another step of re-quantitation, that considers those patterns that have not been assembled into pairs, but that have been identified by the database engine. Since we know the state of the cluster (labeled or not

labeled) it is possible to calculate at which masses the potential partner is expected. If at least three peaks are found where they are supposed to be, the ratio are calculated. It has been noticed, during this work, that this option widely increases the chance to get some outliers (on the other side it increases the number of quantified peptides). Afterward, it is necessary to assemble peptide hits into protein hits, and this is not a trivial step. To work out this step, it has been introduced the idea of protein group: each peptides may belong to more than one protein.

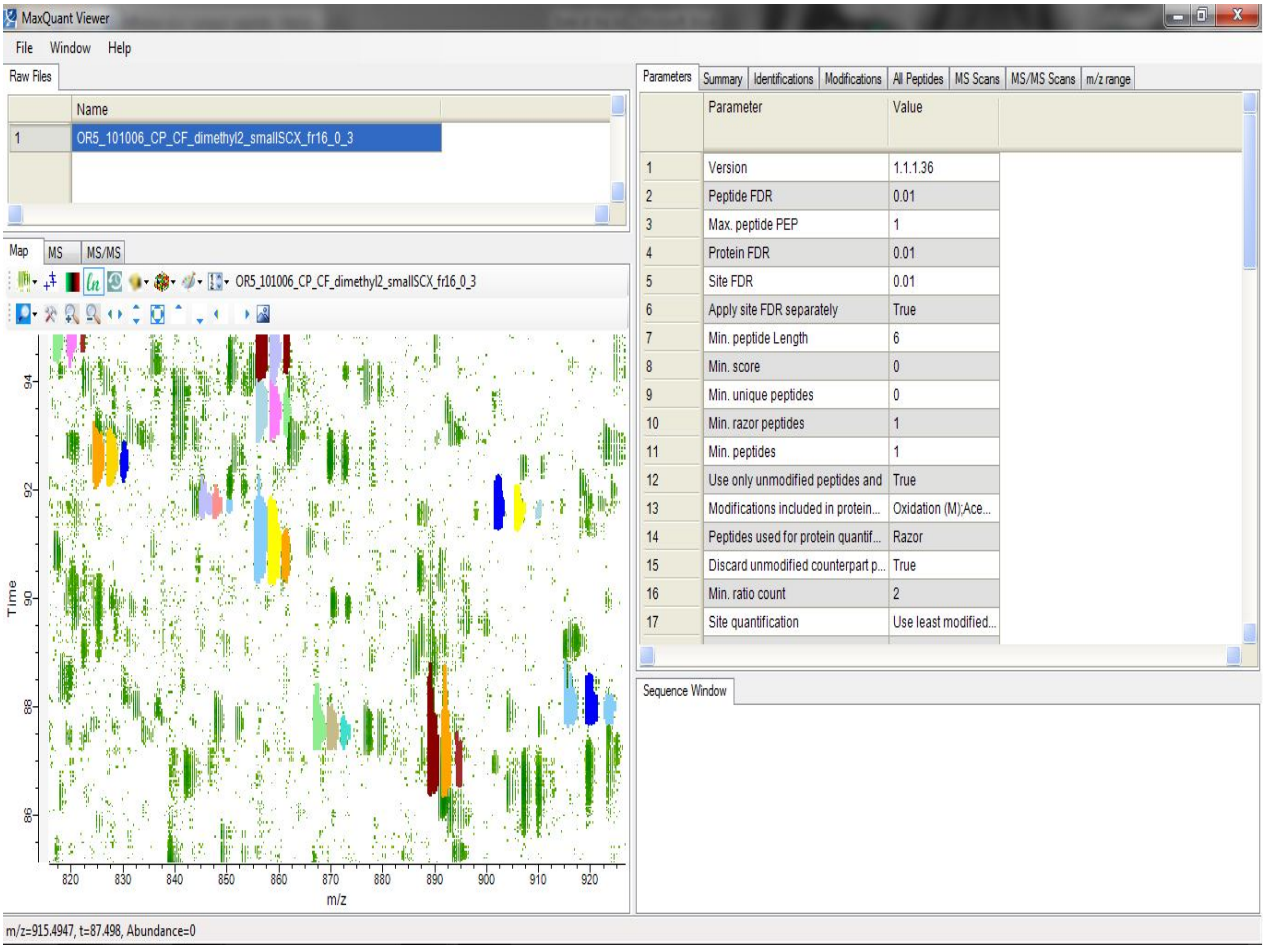


Fig.2: A screenshot from the viewer tool of MaxQuant

To obtain the final quantification it is possible to consider the unique peptides (those which belong only to one protein group) or all the peptides. The protein ratios are then calculated as the median of all the peptide ratios, minimizing the effect of the outliers. Finally are computed two values, called significance A and B,

which are two outlier significance scores, useful for some statistics analysis of the obtained results.

The last part of the workflow is only about the visualization of the achievements. In particular, the viewer tool allows to visualize the data both visually and with some tables.

As shown in the figure 2, the visualization of some part of the bi-dimensional map is displayed on the left corner, where it is possible to point up the multiplets and the isotope clusters with a color-code. It is furthermore allowed to visualize single MS spectra, picked up from the menu to the right of the screen. Moreover, it is possible to view some data on the opposite part of the screen: more specifically the menu at the top allow the user to check out all the MaxQuant results: from the peptides found to the proteins, from the MS visualization to the summary of the parameters used during the quantification process. It is even possible to load more than one raw data, to let the user a matching analysis with different kind of data.

Proteome Discoverer

Proteome Discoverer is a commercial, comprehensive and expandable software platform realized by the Thermo Scientific group. The program is similar to MaxQuant, because it is able to perform both the identification process and the quantitation process. In particular, the multiple database search provides the possibility to combine different algorithms (Sequest, Mascot..) and then maximize and cross-validate the results obtained. Such as MaxQuant, it supports different kind of dissociation techniques and different kind of tagging, like TMT,SILAC and iTRAQ; it even provides False Discovery Rate for the determination of the proteins and the automated annotation of identified proteins with GO classifications and Post Transcriptional Modifications. In the figures 3 and 4 are shown two screenshots of the program (fig. 3 and fig.4).

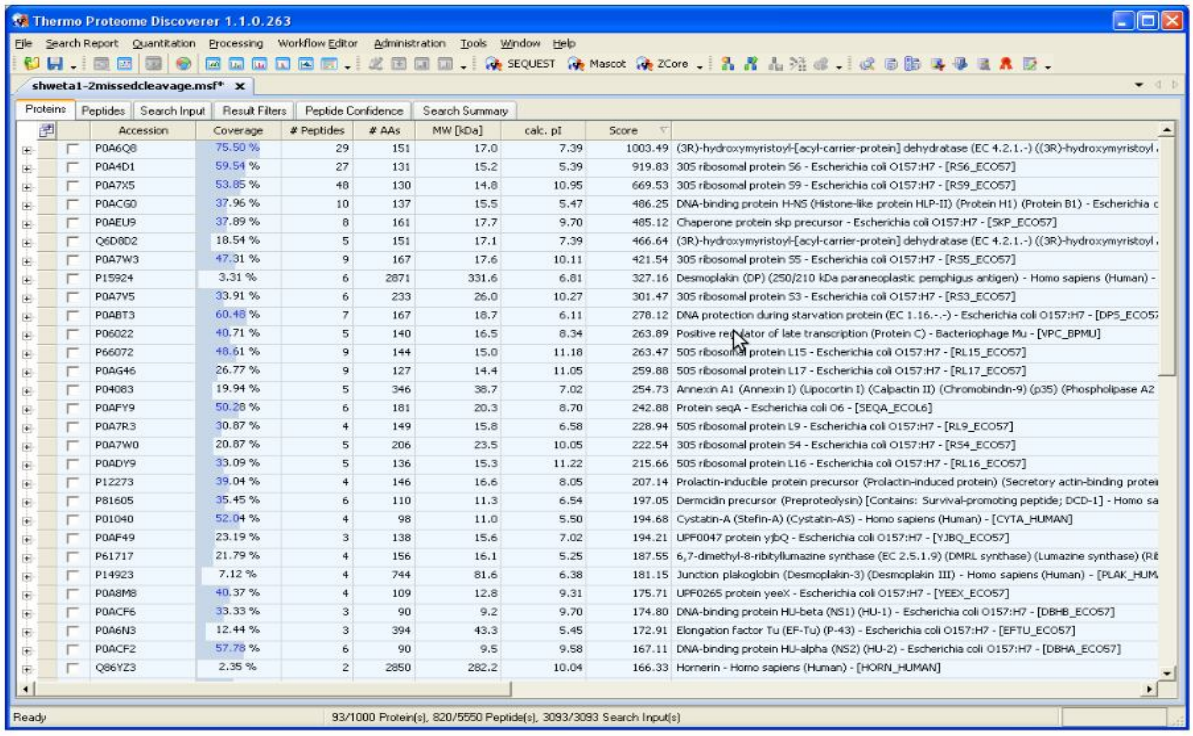


Fig.3: Proteins found by Proteome Discoverer in a sample

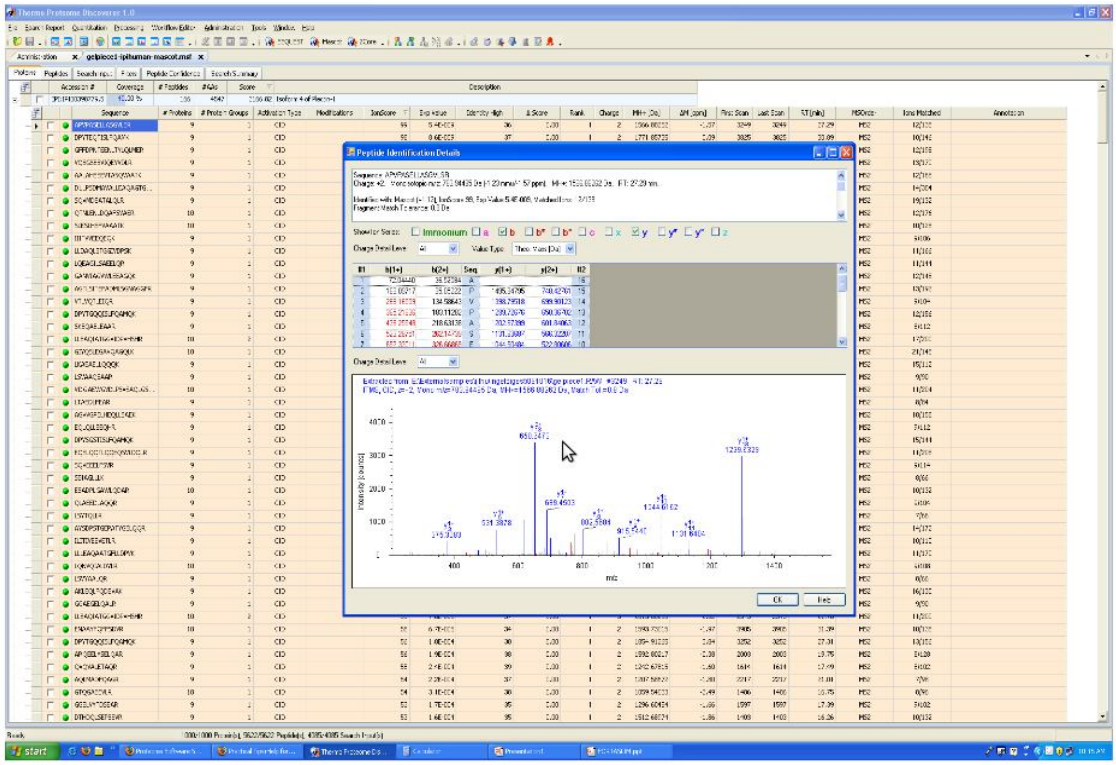


Fig.4: Proteome Discoverer Menu with peptides and a Scan Viewer

Since Proteome Discoverer is a commercial tool, there isn't any documentation about the algorithm used to perform its task, and therefore it is not possible to compare its method with those of MaxQuant. In a paper (which has just been submitted for *Nature* by M. Altelaar, C. Frese et al. from Utrecht University [17]) about the benchmarking between different kind of labeling, it has been performed a comparison between MaxQuant and Proteome Discoverer, showing a little difference between them in terms of results in the quantification process. In the next paragraphs, we will compare our results both with MaxQuant and with Proteome Discoverer, highlighting the differences between them in terms of average and standard deviation of the results obtained at a peptide level.

Methods

In this chapter the operative part of the work will be introduced. In particular the main ideas at the basis of the algorithm will be shown, and how it has been implemented in two different versions, in order to tackle different problems revealed in the datasets used. In particular, the methods proposed aim to fix the problem related to the overlap issue. The overlap happens when two features (a feature is the elution areas of the peptide, spread both in the retention time and in the m/z axes) occupy the same area: the final result is a signal given by the summon of the overlapped distributions. This important trouble may happen in two different situations, with the same effects: the final ratio of the peptide is irremediably poorly quantified. In one case, the overlap happens between two features of the same peptide: in particular, when the shift due to the labeling isn't long enough, then the last peaks of the first distribution (i.e. the Light) are overlapped with the first ones of the second distribution (i.e. Medium). This obviously causes a modification of the final quantification ratio, because the intensity of one feature is enhanced by the contribution of the other one. Another case of overlap happens when two different peptides co-elute in the same area, partially or totally. This situation may happen when the biological sample is very crowded, and there are several peptides with approximately the same retention time and mass-to-charge ratio. To solve this problem, the peaks have been split into several parts along the retention time and the m/z axes, in order to compute different ratios: by comparing them it was possible to exclude those portions which are overlapped. Another idea to increase the accuracy and the precision of the computed ratio, is to use the Pearson Coefficient both in the retention time and in the m/z domain, in order to evaluate those peaks which are properly shaped. In particular, the Pearson coefficient has been used along the m/z axis to discard the noisy isotopic distributions, as we are going to see in the remainder of this chapter, while the correlation along the retention time axis is used to score the quality of the

quantification performed.

Hence, in the next paragraphs we are going to descrive in details these ideas and how they have been implemented in the Matlab framework. The first part is dedicated to the main ideas used in the algorithm, in order to understand how they work. In the second part of the chapter, instead, it is described in detail the workflow of each algorithm, and how the ideas previously described are used.

Quantifying algorithm: the concepts

To properly understand the workflows and the ideas on which the algorithm is based, it is important to recall the basic concept that each peptide feature has a complex three-dimensional morphology, generated by the simultaneous elution of all its isotopic components in the LC-MS map (fig.1).

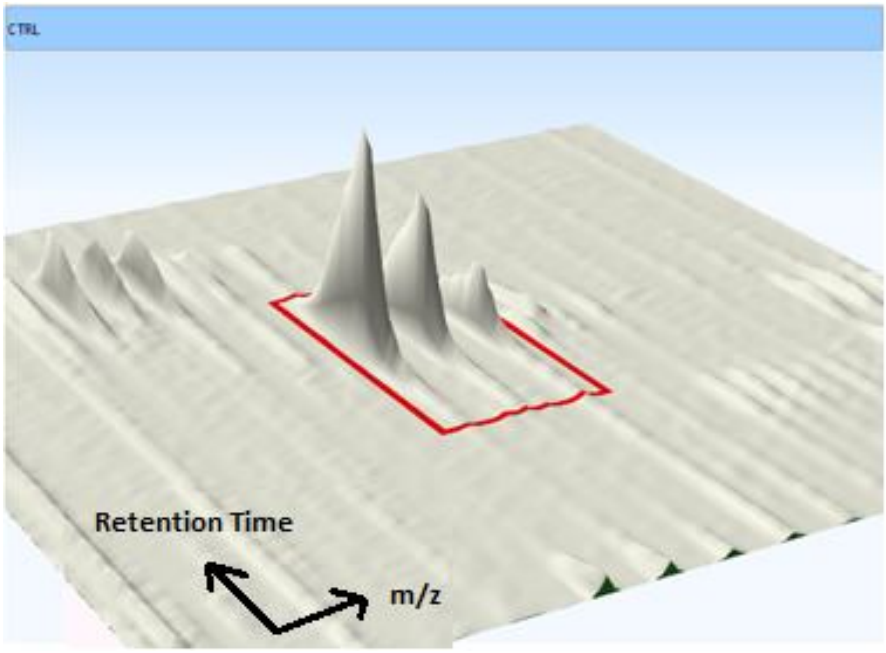


Fig.1 Example of features in the Retention Time – m/z domain

The length of the isotopic distribution in the m/z axis is related to the number of peaks whose intensity is higher than the noise, (typically three or four peaks), while in the retention time domain the length of the elution profile is intensity-dependent: this means that the higher is the signal, the longer will be the profile.

The shape of the profile along the retention time is also related to the chromatography process upstream (see the introduction chapter): usually, an Exponentially Modified Gaussian is used as a model, because often the last tail of the Gaussian is longer than the first one (due to the interaction with the chromatography column – fig.2). Therefore, it is possible to operate on the signal both in the m/z and in the retention time domain: in particular, we have applied the Pearson coefficient in both directions, to check the similarity of the elution profile between the different isotopic peaks of the same peptide and to evaluate the matching of its isotopic distribution to the theoretical one; in this way it is possible to use only the uncorrupted information to properly quantify, as shown in the next paragraph.

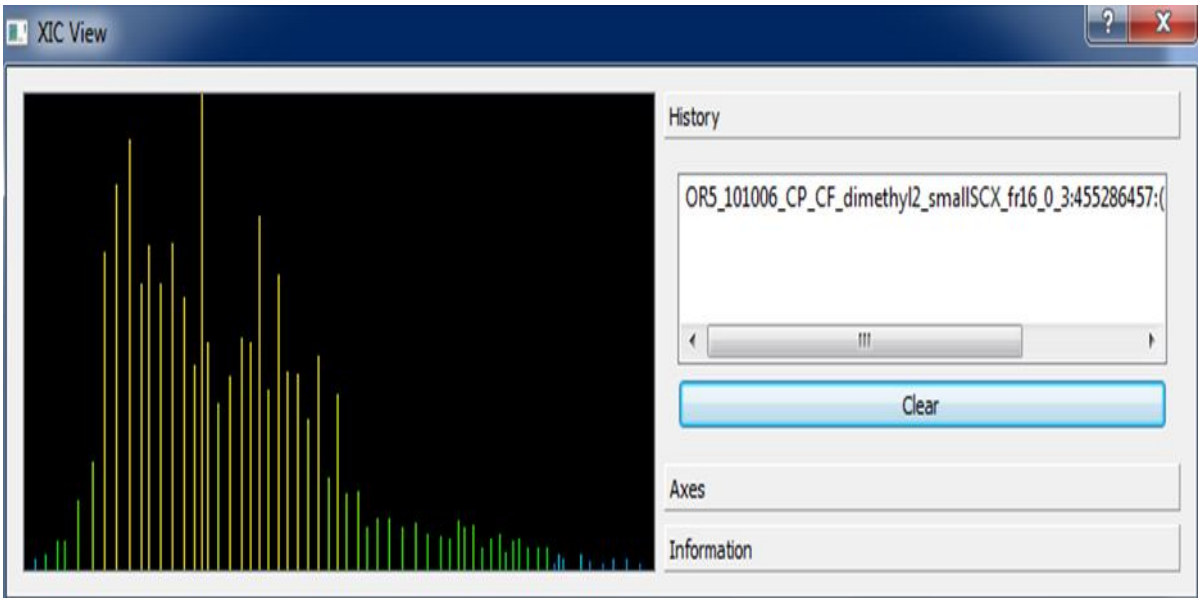


Fig.2 An example of Extracted Ion Chromatogram: the elution profile of a peptide

Pearson Coefficient along the Retention Time axis

For every peptide identified, we can consider at least three peaks for the isotopic distribution labeled as ‘Light’, three peaks for the one labeled as ‘Medium’, and other three peaks for the ‘Heavy’ one. When we use the term ‘peak’, we refer to the three-dimensional Gaussian located at a specific m/z value, and with an elution profile in the retention time. Therefore, for each feature we have at least

nine peaks available to compute the quantitation ratio. It is possible to compare such peaks, in order to understand if some of them is different in the elution shape: the difference in the retention time shape is a clue of a possible overlap with another peptide.

The coefficient is obtained as a ratio between the covariance of the distributions and the product of their standard deviations:

$$\rho_{x,y} = corr(x,y) = \frac{cov\,(x,y)}{\sigma_x \sigma_y} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$

The meaning of such coefficient is the linear correlation between the two profiles. The value ranges from the perfect and positive correlation equal to 1, to the perfect negative correlation equals to -1. If the value is equal to 0, then the distributions are linearly not dependent. For our purposes, the linearity matter is not a limitation, because theoretically the elution profile for each peak should be very close with the others in terms of trend, because all of them belong to the same peptide and there should be only variations in terms of intensity. As already said, it may happen, that isobaric peptides with the exactly the same mass-to-charge (m/z) ratio elutes almost at the same retention time, and therefore it partially overlaps with the peak under investigation, as shown in the figure 3 (the co-eluting peptide problem). When it happens, the Pearson correlation between the three peaks of the same isotopic distribution points out a difference in the shape of the elution profile, and then it is possible to tackle this overlap in different ways.

The first idea should be the elimination of the whole peak from the quantitation process, but it may be a problem because we would lose a big amount of information useful for our purposes. It would be useful to deconvolute the overlapped peaks, in order to separate the contributions of the overlapped peaks, but doing it analytically may be not feasible due to the noise, then we have approached this problem splitting every peak in different parts, and comparing them (see next paragraph).

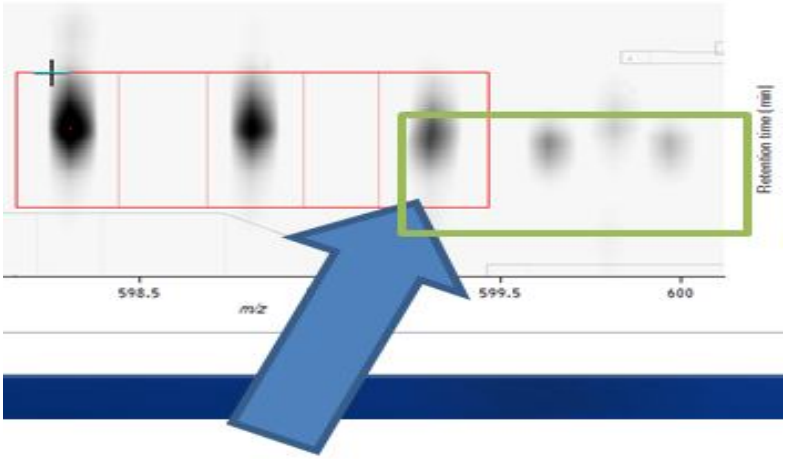


Fig.3 Two features partially overlapped

The main reason to compute the Pearson coefficient along the retention time domain is to use it during the scoring process, in order to be able to select only the best rated results.

Pearson Coefficient along the m/z axis

As for the retention time, it is possible to compute the Pearson coefficient also in the m/z axis. The main difference is that we are going to calculate the Pearson value only on a three-long signal, which is the isotopic distribution composed by its three first peaks (those which are relevant for the quantitation intent). Another important difference between the application of the Pearson coefficient in the m/z domain, rather than in the retention time domain, is that we are going to compare the isotopic distribution of the real signal with the theoretical model of the distribution, while in the retention time domain we compare the peaks between themselves. This difference is mainly due to the possibility to have the exact (not really exact as we will see, but close enough for our purposes to the real one) isotopic distribution, starting from the mass value of the peptide. To get such distribution, we have exploited a model realized in the 2008 by Valkenborg et al. [19]. This model, which is based on a polynomial model, required as input only the mass of the peptide, which is easily inferred from the m/z value and the charge, and the intensity of the first peak, which is equal to one being normalized. As it is

possible to see from the table 1, the difference between such model and that one from IPC [20] (Isotope Pattern Calculator, which is used as reference and it exploits the exact sequence of amino acids of the peptide to get the distribution) is very close, anyway close enough for our comparing purposes. In the table the averages and the standard deviations of the values obtained as difference between the peaks of the IPC distribution and those of the Valkenborg model, out of 64447 simulated isotopic distribution, are shown. Since the distribution were normalized, the first peak were equal to one in both of them. As it is possible to see, the average difference is quite small.

	Peak 2	Peak 3	Peak 4
Mean Error	0.0242	0.0353	0.0527
Std Error	0.0316	0.1092	0.0951

Tab.1 Valkenborg model evaluation

If the Pearson coefficient computed along the retention time axis was used only for scoring purposes, the Pearson coefficient computed along the m/z axis is useful also for the quantitation process. Indeed, each scan across the elution area of each peptide's feature is used to compute the correlation with the theoretical distribution: if the value isn't higher than a fixed threshold, then the scan is discarded and not used for the final quantitation.

Taking Advantage of Several Ratios

As already stated before, the main idea is to split the elution area of the peptide in different parts, three or five, as shown in the figure 4. Each part is composed by several isotopic distributions. We can then select the scan that pass the Pearson selection along the m/z axis, and average their isotopic distributions, obtaining the isotopic distribution of the first tail, that one of the body of the Gaussian and finally the distribution of the last tail.

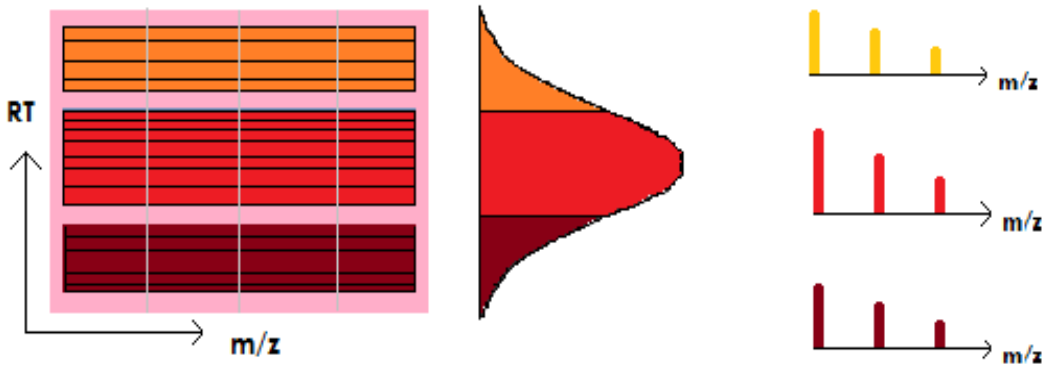


Fig.4 Division of the elution area in three parts and their averaged isotopic distributions

In this way, it is possible to have three different distributions for the same elution area. Since we have two elution areas for each feature (the one of the Light and that one of the Medium - or Heavy), we can compute three different ratios, matching coherently the distribution of the first part of the Light with the first part of the Medium (Heavy), the central part of the Light with that one of the Medium (Heavy) and finally the last tail of the Light with that of the Medium (Heavy). It's very important to point out that, in this way, it is possible to compute the ratios with the same parts of the Gaussian, without mixing different areas. Let's now see what happen in case of an external overlap, as shown in the figure 5.

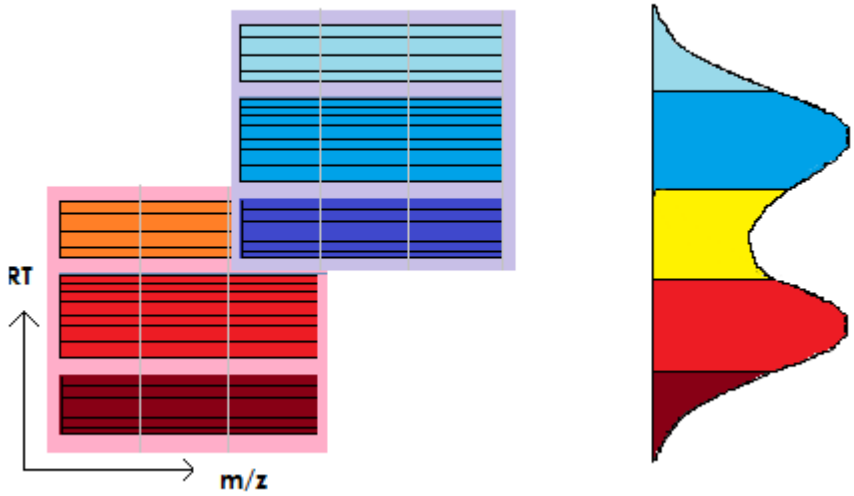


Fig.5 Co-eluting Peptides overlap

It is possible to see that the computation of the final ratio would be affected by the partial overlap of another peptide with approximately the same retention time and m/z value. If we computed three different ratios, instead, it is possible to notice that the co-eluting area would affect only one ratio out of three (if we consider our peptide as the red one, the only part affected by the overlap would be the orange one). Therefore, comparing the three ratios, one of them (the overlapped one) would be different compared to the others, and therefore not considered for the computation of the final ratio. The exclusion of the 'affected' area will increase the accuracy of the quantitation. In the first dataset used, where the label was the Dimethyl, the partition of the area has been made with three different sections. In the other datasets, with the SILAC labeling, the area has been divided in five different parts, because we knew that those datasets were very crowded (therefore the co-eluting problem would be very relevant), and increasing the number of sub-areas the accuracy of the results should increase. To choose which areas using to compute the final ratio, we computed the differences between the ratios of each area, and then they are clustered together in groups, as explained in the paragraph about the workflow. Once we get the clusters (there may be one, or two clusters at most), we compute the final ratio with the longest cluster, or with that one which have the closest ratios. In the vast majority of the cases, there is only one group of ratios, composed by more than the half of the sections.

Multi charged peptides

Finally, the last idea implemented in our algorithm is the exploitation of the multi charged peptides. In particular, it is possible and probable that a peptide, during the ionization step in the mass spectrometer workflow (see the introduction chapter for details) is ionized in more than on charge state. Therefore, it is possible to find the same peptide, and theoretically the same information about the quantitation of that peptide, in two different positions of the bi-dimensional map; it is important to recall that the x-axis of the signal is the mass/charge ratio, and therefore varying the charge of the peptide will vary even the position along the

m/z axis, but not the position in the retention time, which should be the same. Having the same information in two different places, it is possible to exploit such advantage. In particular, when this situation happens, we perform the quantitation process for both of them, but we get as a final ratio for that peptide, only the one which has higher score. Therefore, if one peptide presents an overlap, and its score is low (as the Pearson coefficient detects such issue) we will get the quantitation information from the same peptide with different charge, which probably doesn't have an overlap problem. In this way, it is possible to increase the accuracy and the performance of the quantitation process of each peptide.

So far we have seen the main ideas at the basis of the algorithm. Let's now see in detail how it works.

Implementation and Workflow

In this second part of the chapter, the workflow is described in detail. Moreover, the different ways to quantify are shown, such as the method proposed by Yoon to solve the overlap issue between the labeled features of the same peptide, and the partitioning of the elution area in several sections.

The overlapping issue

This algorithm may be ideally divided in five parts, as shown in figure 6. The only inputs required, such as the first algorithm, are the mzXML file (readable in the Matlab environment), and the output file (read in Matlab as an excel file) provided from the software used for the identification of the peptides (MaxQuant or Proteome Discoverer). To get the mzXML files for each dataset from the raw file coming directly from the mass spectrometer, we have used a function named '*msconvert'* provided by the '*ProteoWizard'* library, freely available on the web. As already said, the identification of the peptides is performed upstream by the software used for the identification process, either MaxQuant or Proteome

Discoverer. Starting from the given retention time, we have to find out the begin and the end of the peptide along the retention time axis. MaxQuant provides at least the length of the peptide in the retention time (but not when it starts or finishes, just its length); but Proteome Discoverer doesn't. Therefore we have implemented two different approaches for each program used, in order to localize the peptide's elution area. Once the elution area has been localized, it is possible to proceed with the quantification process.

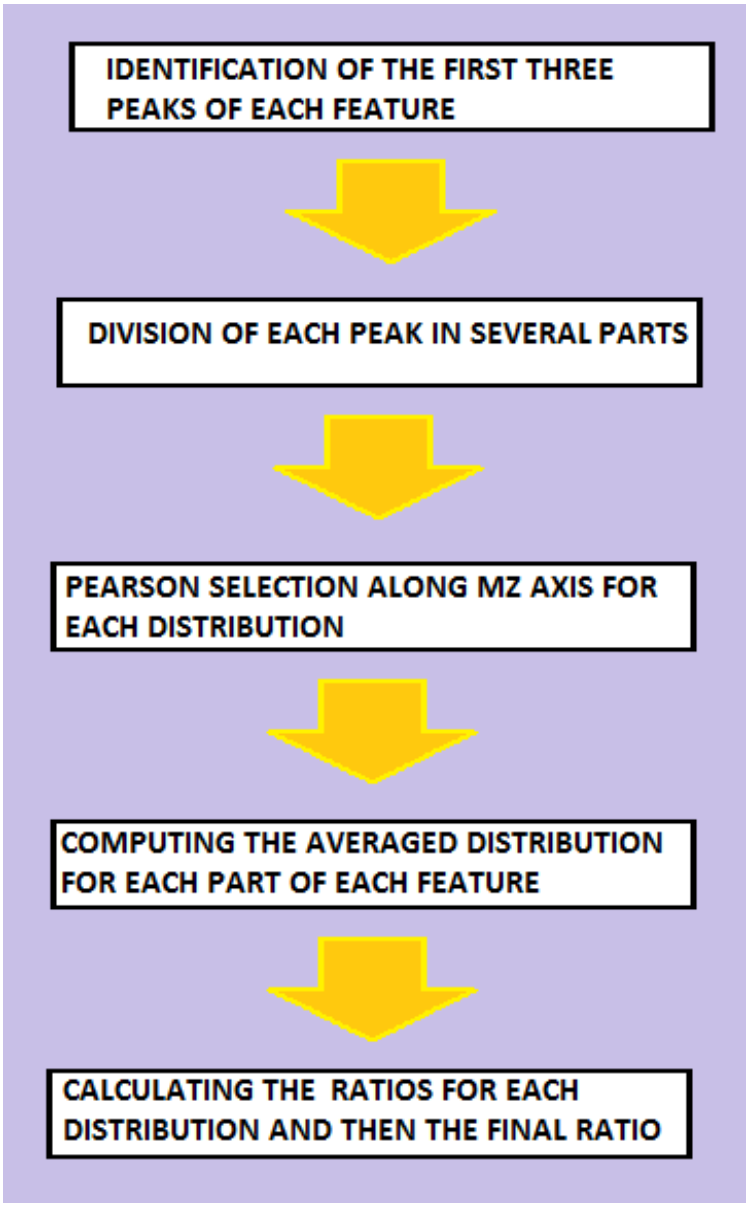


Fig.6: the workflow of the algorithm

The algorithm can be ideally divided in two workflows, really close each other but substantially different in the quantitation strategy. The first one is optimize to solve the overlap between the features of the same peptide, and it is based on the implementation of the Yoon's method (see next paragraph for details); the second one is basically the same, but the features are not overlapped, and the quantitation may be done easily. Moreover, in the second algorithm the idea of scoring the quantification process has been introduced, and to solve the problem related to the co-eluting problem, the elution area is split in five sections rather than three.

The overlap between features of the same peptide happens because in the peptide sequence there is no *Lysine* amino acid, that is the binding site for the labeling. As already stated in the overview, the digestion of the proteins is usually performed enzymatically, by the Trypsin, which cuts the sequence where there is a Lysine (notated as 'K') or an Arginine (notated as 'R'). When there is no Lysine in the sequence of the peptide, the label is attached only in the N-terminal part of the peptide, causing a shift of only four Dalton (this is the length of the shift caused by a Dimethyl labeling; for the other dataset used, where the labeling is a SILAC, there isn't any relevant overlap between the features). This means that the fifth peak of the Light distribution is overlapped with the first peak of the Medium distribution, and obviously the fifth peak of the Medium is overlapped with the first of the Heavy. As said, this causes an alteration in the final ratio. For this reason, it is necessary to compute separately the ratios of the overlapped peptides and the ratios of the not overlapped peptides. In the next figures (fig.7 and fig.8) the two versions of the algorithm are shown: the first one is specifically designed to tackle those peptides with an overlap problem. It is possible to see from the pictures that the workflow is very similar in the two cases. The main difference is related to the quantification method, which is explained in the next paragraph. Once that the elution area has been identified (1), it is possible to divide such area in three parts (2). In this way, at the end of the process we will have three ratios to compare, in order to get the final ratios, as explained previously (*several ratios*). Before computing the final isotopic distribution for each area, calculated as an average of

the distributions from each scan, it is performed a selection based on the Pearson coefficient (3).

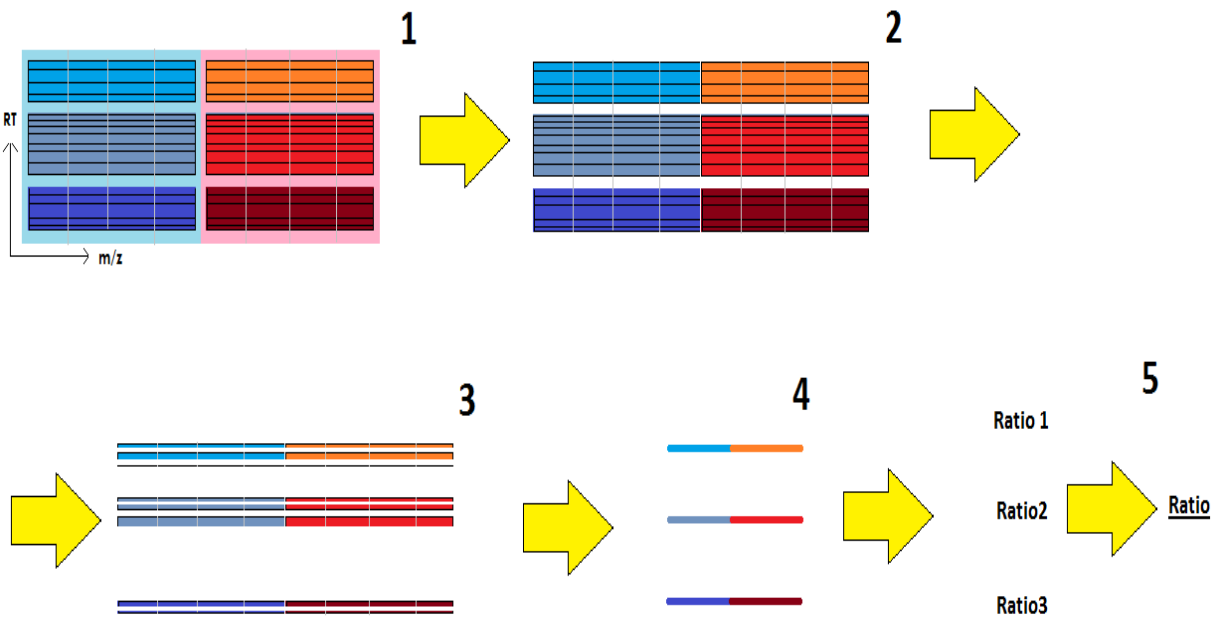


Fig.7 Workflow of the algorithm for overlapped peptides

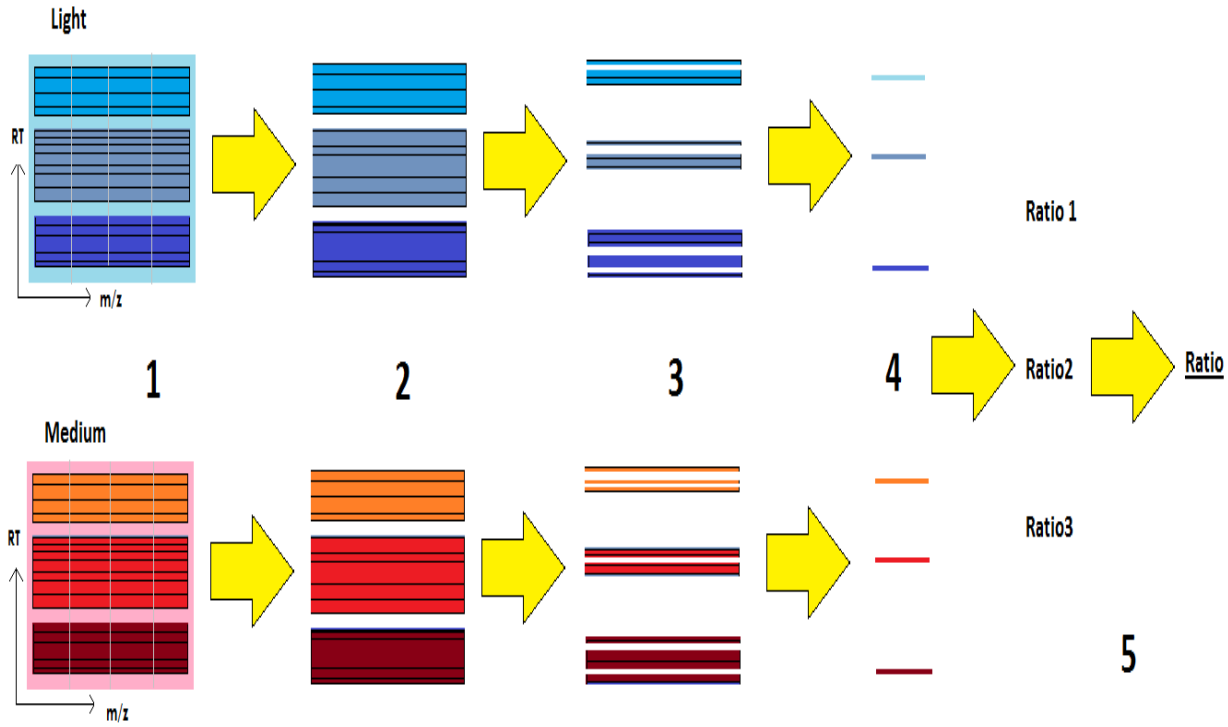


Fig.8 Workflow of the algorithm for not overlapped peptides

It is important to notice that the Pearson correlation is performed only along the m/z axis, using the Valkenborg model to get the theoretical distribution. Therefore, only those scan whose correlation coefficient is higher than a threshold are used in order to compute the averaged isotopic distribution for that area. At the end, three different ratios are obtained (4,5). How these ratios are calculated is described in the next paragraph. It is important to recall that these ratios are compared, and if the difference between them is smaller than a threshold, they are used to get the real final ratio. To do so, the ratios are sorted, and it is checked the difference between the first ratio and the second, and between the second and the third. For the final computation are kept only those ratios whose difference was below the threshold. Before explaining how it is performed the quantitation, it is interesting to point out that the algorithm is basically based on two tunable threshold: the first one is about the Pearson correlation in the m/z domain, the second threshold is about the difference between the ratios. In the results, we are going to see how tuning such parameters, the results will vary.

Yoon's method and the linear coefficient

As already said, in the algorithm it is possible to quantify in two different ways, on the basis of the presence or absence of the Lysine amino acid. In the case of absence of the Lysine, we have to tackle the overlap problem between features of the same peptide. To do it, we have used a method implemented by a Korean researcher in 2010. [18]

The algorithm proposed, as stated in the paper, can be even easily applied in the Trans Proteomic Pipeline (TPP, it is a widely-used freely available proteomics pipeline), during the peptide quantification, because of its simplicity, and it uses the bi-quadratic equations to properly solve the overlap. Before understanding how these equations may be useful, it is necessary to figure out the model of the two clusters overlapping; for this reason, figure 9 may be helpful.

L1	L2	L3	L4	L5	L6				
				H1	H2	H3	H4	H5	H6
I1	I2	I3	I4	I5	I6	I7	I8	I9	I10

Fig.9: the Yoon model of the overlap

The first pattern (labeled as light – 'L') is composed by six peaks - to simplify the discussion without losing generality- but it may be even longer; the same is for the second cluster (labeled as Heavy – 'H'). The available data from the raw file are named 'I', and they are obviously the sum of the two distributions. It's quite easy to agree to this mathematical formulation, which formalizes the model:

$I_k = L_k$ if k<=4
$I_k = L_k + H_{k-4}$ if 4<k<=n
$I_k = H_{k-4}$ if k>n

At this point, Yoon states that the sought ratio is:

$\alpha = H_k / L_k$

Then, he deduces a quadratic equation

$I_1 \alpha^2 - I_5 \alpha + I_9 = 0$

starting from two different equations:

$\alpha L_5 = I_9$

$I_5 = L_5 + \alpha L_1$

Using the quadratic formula it is possible to obtain two values for α :

$$\alpha = \frac{-I5 \pm \sqrt{I5^2 - 4 * I1 * I9}}{2 * I1}$$

The sign 'plus' is chosen if the solution is greater than L_5/ L_1, otherwise it is chosen the sign 'minus'. As seen, this quite easy solution allows to find out the ratio with some simple algebraic steps, which are very fast from a computational point of view. Thanks to the simplicity and elegance of this rigorously mathematical approach, we are able to obtain the quantitation of a peptide, even if there is a troublesome overlap between the labeled and unlabeled pattern.

Fortunately, if there is at least one Lysine (the site where the label is bound) in the peptide sequence, then the overlap doesn't happen at all. In that case, the Yoon's

approach obviously is not necessary. To quantify these not overlapping peptides we compute the linear coefficient of the straight line obtained by interpolating the points, which have as abscissa the peak intensities of the Light and as ordinate the peak intensities of the Medium (this is a standard procedure used, for example, by MaxQuant).

The division in several parts

In the dataset where the overlap between features of the same peptide doesn't occur, due to the SILAC labeling, then, the workflow is slightly different. In particular, it is given much more importance to the co-eluting overlap. Thus, once the peptide is localized, it is possible to accomplish the quantification process. The workflow is schematically shown in the next figure (fig.10).

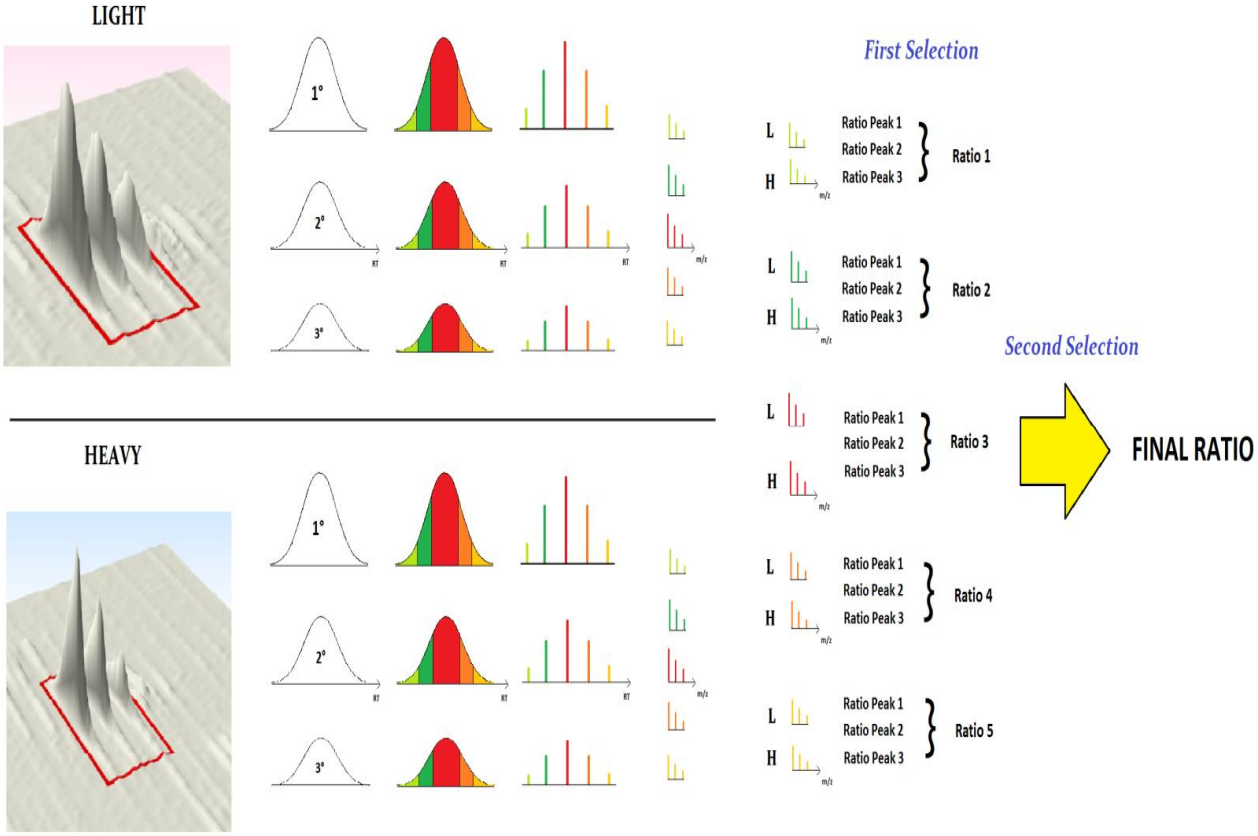


Fig.10: Schematic representation of the workflow

It is possible to see that we have globally six peaks: three peaks come from the Light distribution, and three peaks come from the Medium distribution (or Heavy, it depends which ratio we want to compute at the end of the algorithm). The next step requires the elution profile of each peak to be divided in several parts (it has been set to five); these parts are supposed to be equally long. Hence, we can consider each part and compute the isotopic distributions for every scan: for all of them, we select only those whose Pearson correlation with the theoretical model is higher than a fixed threshold. Finally, it is possible to reconstruct the isotopic distribution along the m/z domain, averaging all the distributions which have passed the Pearson selection: we have then five isotopic distributions for each feature. At the end, it is possible to calculate the desired ratio.

To compute the final ratio, at first, we calculate three ratios for each couple of distributions (which come from the same part of the original peak – in the picture they have the same color), one ratio for each couple of peaks. Once we have these three ratios, we compare them, and we compute the ratio for that distribution, averaging only those values which are close enough (whose difference is lower than a threshold). In the picture, this is named as *First Selection*. Doing this procedure for every couple of distributions, we computed the five ratios (one for each couple). Ideally, the five ratios should be very close to each other, because the feature of the Light and that of the Medium should be proportional in each part considered. Unfortunately, this doesn't happen ever, and we have to choose which ratios considered for the calculation of the final ratio (*Second Selection*). First of all, we cluster the five ratios in at most two groups, based on the difference between the ratios. We can even have no cluster at all, if the ratios are far away from each other, but this doesn't happen very often. Usually, there is only one cluster or, at most, two clusters (two ratios per group, or two and three ratios). If we have only one cluster, the final ratio is the average value from that cluster. If we have two clusters, we considered the biggest cluster or, if they have the same length, we choose that cluster whose ratios are closer in percentage. In this way it is possible to compute the final ratio for that peptide.

It is interesting to highlight the roles of the two selections, as shown in the next figure (fig.11). In the upper part of the figure, we can see the effect of the first

selection: if there is an overlap on the third peak, the ratio coming from the third peak will be discarded, compared during the first selection with the ratios coming from the first and second peak. In the lower part of the figure, instead, we can appreciate the effect of the second selection: if there is an overlap on an area of the peaks, let's say the fourth and the fifth part, the ratio coming from that area will be discarded during the second selection, because the first three parts will have different ratios (the corrected ones).

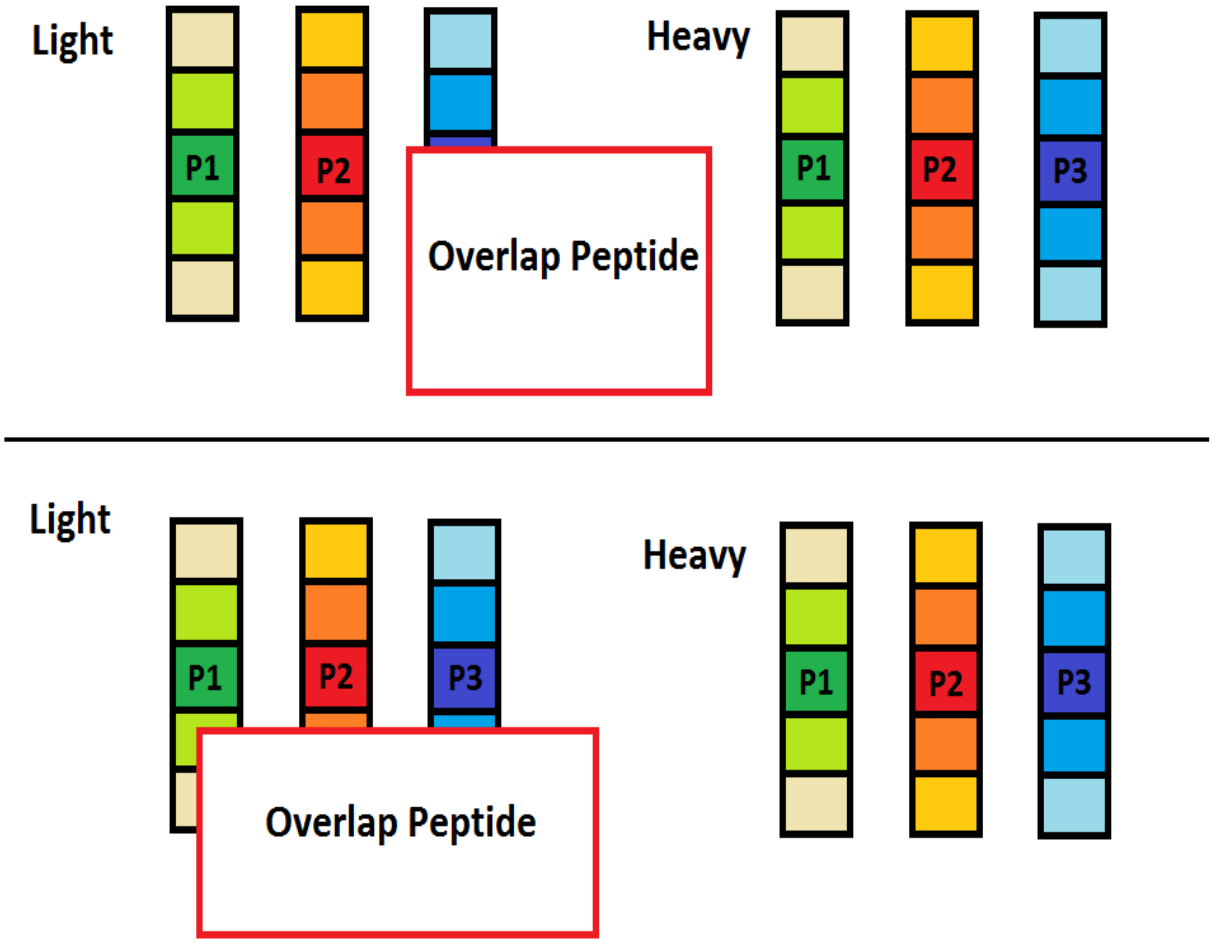


Fig.11: Schematic representation of the overlap

To summarize, it is important to highlight that this first 'selection' has worked considering the peaks from the isotopic distribution: in other words we have

operated a selection along the m/z axis; the second selection, instead, allow us to perform a screening along the retention time axis, covering in this way all the domain of the bi-dimensional signal.

Scoring the quantification process

The scoring process is realized in order to give a rank about the quality of the quantitation for every peptide. As a matter of fact, it is important to be able to measure the probability that the ratio obtained is corrected, and be able to grade the whole quantitation process. Generally, the evaluation of the quantitation is done at a protein level, not for every peptide. In the march 2011 Reiter et al. proposed an algorithm with a peptide ranking method, mainly based on the quality of the peaks used for the quantitation [21]. In our algorithm, there are at least three elements which contribute to the final evaluation of the quantitation algorithm. Such elements are:

1. The Pearson Coefficient obtained along the Retention Time axis
2. The score obtained in the identification process (provided by MaxQuant or Proteome Discoverer)
3. The numbers of sections of the elution area used for the final quantitation

The Pearson coefficient along the retention time axis has been discussed previously. It's interesting to highlight that such value is really important to figure out if the peptide being quantifying is properly shaped, and if there is no overlap with other peptides. The score obtained in the identification process is mainly related with the intensity of the signal from that peptide: if the peptide has an high intensity, it will be properly identified, and it will be easier to exactly quantify. Finally, the numbers of sections used gives us the idea about how many ratios have been considered to compute the final ratio: higher is the number, better is the quantification. Considering all these values, we set up a linear combination of such parameters, properly weighted in order to give an higher score to that peptide which is closer to the expected value. The final formula about the score is:

$$score = \frac{pearsonRT * W1 + scoreId * W2 + numbersOfRatios * W3}{W1 + W2 + W3}$$

The values of the weights may be changed, in order to point out only those peptides with some specific feature, such as a proper shape in retention time, or an high value in the identification process. In the figure 12 it is shown how the peptides are quantified (ratios on the abscissa) in function of their score (on the ordinate), when the values of the weights are all set equal to one.

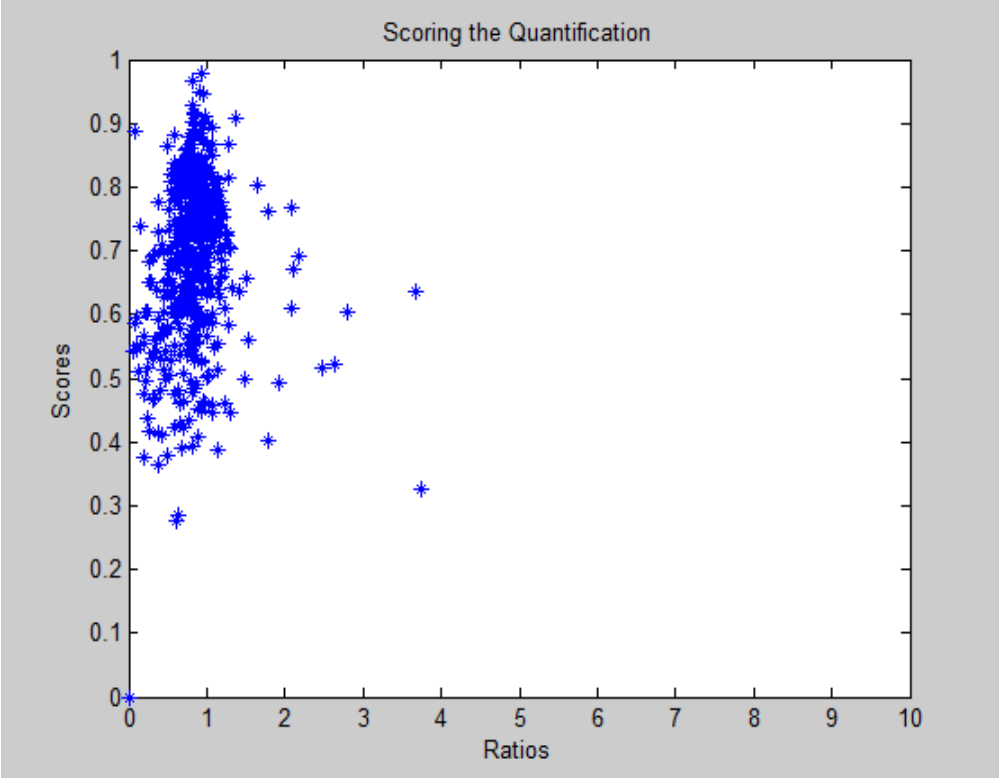


Fig.12: Scoring versus Ratio

To get the best combination of weights, it has been implemented a simple linear classifier, used to select only those peptides that have been quantified properly. After the selection of such peptides, it is possible to see that the quantification is significantly improved. In the next chapter -about the results- this idea is described in detail.

Results

In this last chapter we are going to show the results obtained with the algorithm described in the previous chapter. This last part can be divided in two parts, since our algorithms have been tested on two different kinds of data: one dataset has been realized using the dimethyl labeling, where the overlap between the feature of the same peptide may occur; instead, the other datasets have been realized using the SILAC (SILAC is the acronym for Stable Isotope Labeling by Amino acids in Cell culture) labeling, where the shift is higher and there is no overlap between the feature of the same peptide. In particular we have tested our algorithm on five datasets.

In the first part of the chapter, along the results obtained with the first type of dataset, the importance and the incidence of the problem of the overlap between features of the same peptide will be discussed: we have realized a statistical analysis on 64447 simulated isotopic distributions to assess how the overlap alters the final ratio. In the second part, instead, a comparison will be shown between the results obtained with MaxQuant and those obtained with Proteome Discoverer, on the same raw files. As we will see, Proteome Discoverer (it is important to highlight that this program is commercial, while MaxQuant is not) has higher rate of identification of the peptides, and even better results in the quantification process, in terms of standard deviation of the final results. Our algorithm, finally, increases significantly the accuracy of the quantitation compared with the MaxQuant results, keeping approximately the same number of peptides identified. After a selection based on the score, as shown in the last paragraph, it is possible to further enhance the accuracy of the results, obviously losing some of the peptides quantified.

Dimethyl Dataset

The first dataset used to test our algorithm is the Dimethyl Dataset, realized, as the others, in the Laboratory of the Professor Heck in Utrecht, and it's a fraction of a bigger experiment. In particular, the realization of the mix of proteins has been realized on purpose, knowing exactly that the ratio between the Medium and the Light is equal to 0.5, while the ratio between the Heavy and the Light is equal to 0.1. Unfortunately, this last one value is critical in general for the quantitation task, because it is too low: very often the feature of the Heavy is covered by noise and it is not possible to properly quantify it. Even MaxQuant, the only software used for this dataset, get very noisy results, with a very high standard deviation and not even centered on the expected value. For this reason, we have tested our algorithm only on the first ratio Medium / Light, whose expected value is 0.5. Another important problem, related to the Dimethyl labeling, is due to a shift in the retention time of the features, due to the presence of deuterated heavy labeled peptides, which are known to elute prior to their corresponding light one: it may happen that the Medium isn't exactly aligned with the elution time of the Light, as shown in the figure 1: the Light, which have an m/z value equal to 446, has a retention time higher than the Medium, which is located at a m/z value of 450 Dalton. Another important parameters of the dataset is the enzyme used for the digestion. As usual, in this dataset it has been realized the trypsin digestion, which is the most used due to its several advantages (see the introduction for details). As known, the Trypsin cuts the protein sequence immediately after the amino acids Lysine and Arginine. If the cutting amino acid is the Arginine, it is very likely that there is no Lysine in the sequence, and the peptide is surely affected by overlap between Light and Medium. It may happen that the Trypsin misses some Arginine or Lysine, and therefore the label is doubly attached to the peptide, which will have a longer shift between the features.

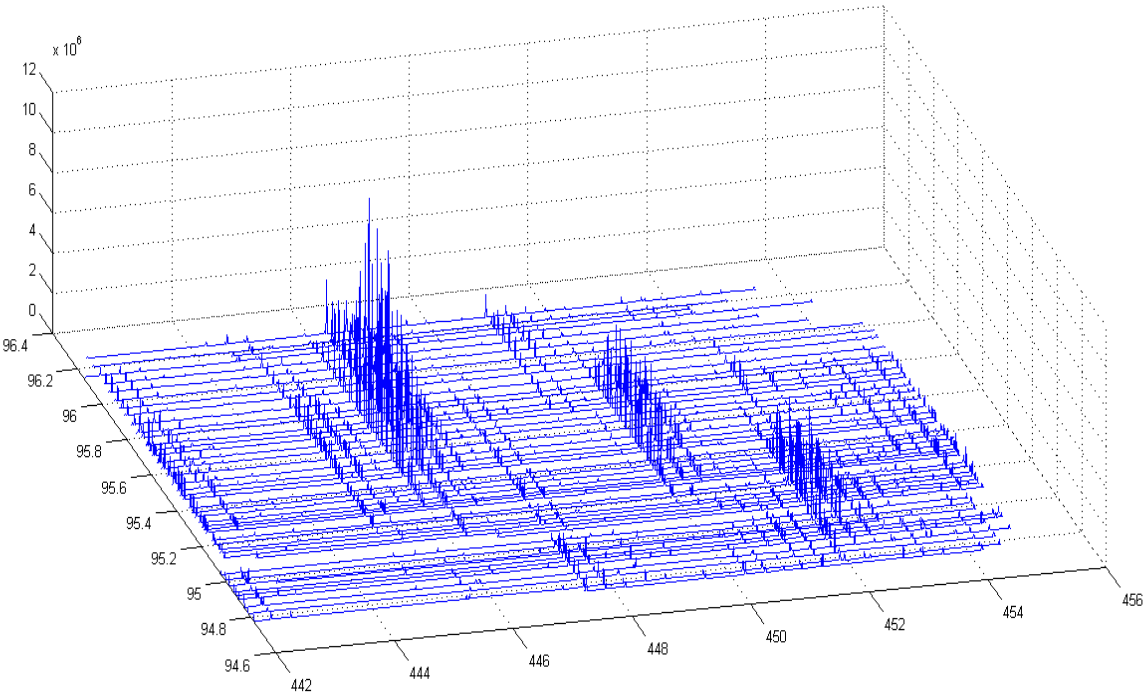


Figure 1: dimethyl retention time shift between feature of the same peptide

The maximum number of missed cleavages generally considered are two for each peptide. Moreover, not only the Lysine is the site for the labeling, but even the N-terminal part of the protein, which is obviously always present. Therefore, the smallest shift between feature is equal to 4 Dalton (as known, the Dimethyl causes a shift of 4 Dalton for each label attached to the peptide).

This Dataset is composed by 31167 scan, both from the first level and from the second level (the scan from the second level are about the fragmentation of the highest peaks of the first level, and they are used for the identification process). In particular, there are 6386 scan from the first level, and 24781 scan from the second level. The range in retention time goes from 20 minutes to 180 minutes, while the m/z range is from 380 up to 1400 Dalton.

Let's now see how MaxQuant is able to perform the identification and the quantitation process, and how it handles the information provides by all these scans.

MaxQuant Results

MaxQuant, during the identification process, has been able to find out 6127 peptides, located in 9743 evidences. The term ‘evidence’, used in the MaxQuant framework, means:

1. A single identified feature of the triplet, standing alone (MaxQuant has been able to identify, for example, the Light distribution of the peptide, but can’t find the others components of the triplet);
2. The whole triplet of the peptide, composed by a Light, Medium and Heavy distributions, properly identified by MaxQuant.

The evidences may be divided (fig.2) based on the presence of the Lysine in the sequence, but even if the triplet of the evidence is complete (it is named MULTI) or if there is only one distribution identified (it is named ISO).

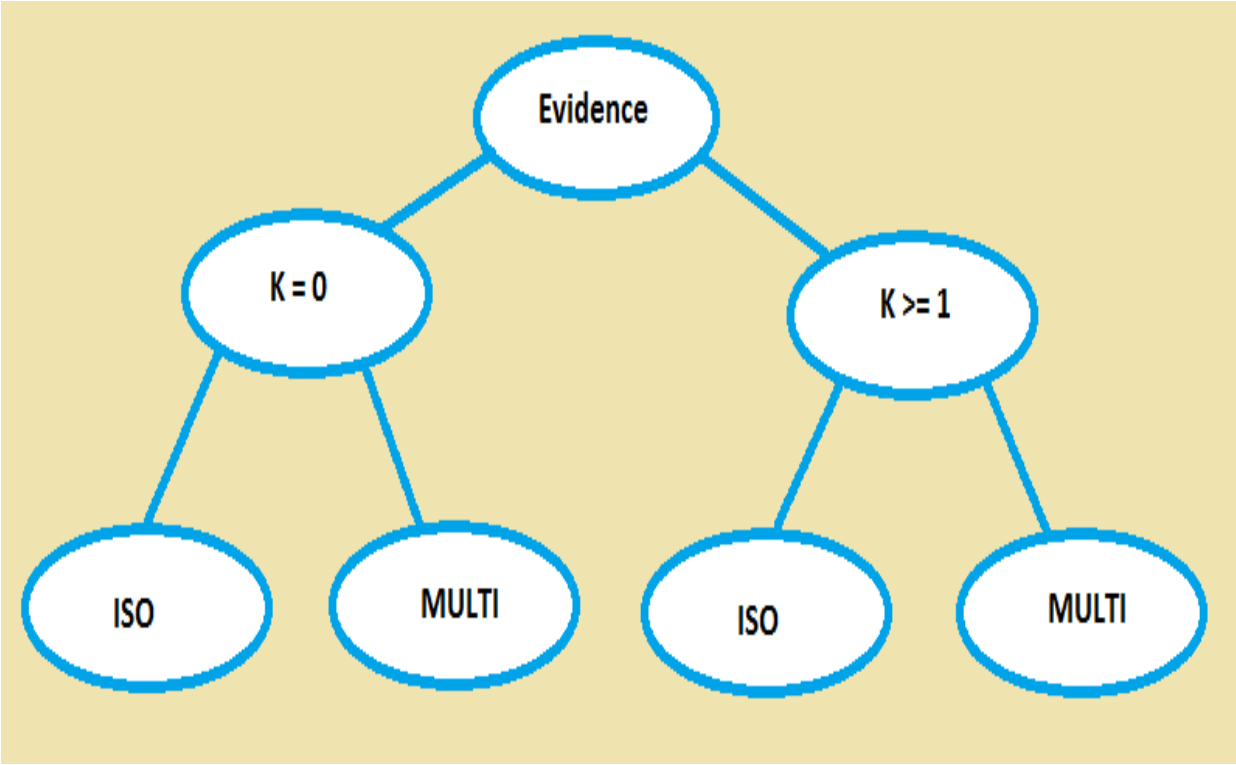
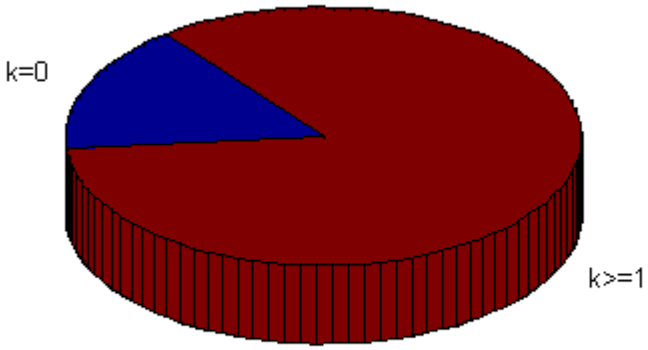


Figure 2: division of the Evidences of MaxQuant

For our purposes the presence or absence of the Lysine is very important, and we will split the results on such basis (even because the quantification strategy, as

seen in the previous chapter, is different: if there is an overlap we will use the specific Yoon's method). Let's see then how many peptides don't have the Lysine in their sequence, looking at the figure 3.

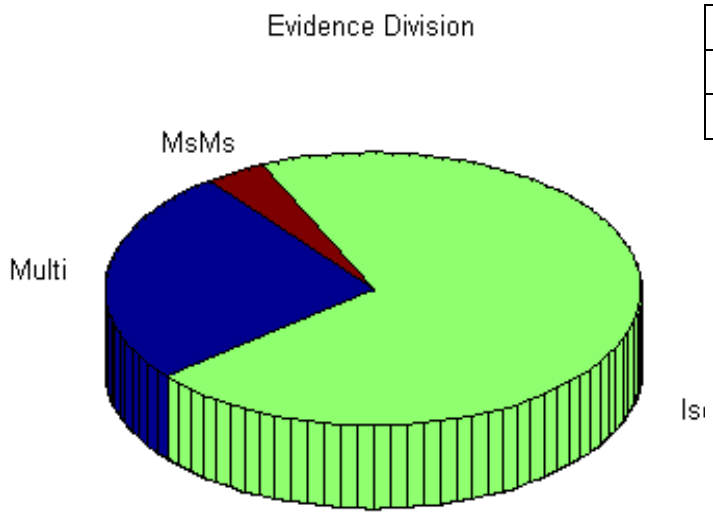


K=0	16.17%
K>=1	83.83%

Figure 3: Peptide without Lysine (K=0) and those with at least one Lysine (K>=1)

It's easy to see that the percentage of such peptides isn't so high, but it is considerably significant, and as we will see in the following, the quantification is a little bit worse than the peptides with at least one Lysine.

The other important division within the evidence file, is between the MULTI and the ISO peptides. It is interesting to notice that such division, as well as that one based on the presence of the Lysine, is not balanced: there are much more ISO distributions than completed triplet (fig.4). There is even another tiny group, named MsMs, which is composed by those peaks which don't belong to any recognized distribution; fortunately, being useless for the identification and the quantitation process, they are very few.



MULTI	25.57%
ISO	70.87%
MsMs	3.56%

Figure 4: division of the Evidences of MaxQuant

Let's now see how does MaxQuant perform the quantification, in particular we will focus on the differences between the results obtained with the peptides without Lysine, and those with at least one Lysine: as we will see, the issue of the overlap affects the quantification decreasing accuracy and precision (and therefore increasing the overall standard deviation). In the next table it is shown the MaxQuant results about those peptides which have no Lysine at all (tab.1). In the *Reference* column there are shown the expected values: as known the Medium / Light ratio should be equal to 0.5, and MaxQuant gets a result which is pretty close to the expected value.

K=0	MaxQuant Results	Reference
Average	0.5503	0.5000
Standard Deviation	1.4493	0
Numbers of Peptides	1530	1576

Tab.1: MaxQuant results: peptides without Lysine

Obviously the standard deviation should be as close as possible to zero. Finally, the last row, named *numbers of peptides*, shows the number of quantified peptides (*MQ* column) out of the identified peptides (*Reference* column).
In the next table, similar to the first one, there are shown the results obtained for those peptides with at least one Lysine.

K>=1	MaxQuant Results	Reference
Average	0.5357	0.5000
Standard Deviation	1.2063	0
Numbers of Peptides	7869	8166

Table 2: MaxQuant results: peptides without Lysine

It is possible to see that, even if the number of peptides in this second table is much more bigger than the first table (7869 versus 1530) the standard deviation and the average value is better in this second table. The average value is closer to the expected value, and the standard deviation is even smaller, (even if the number of quantified peptide is five times bigger!). This, without any doubt, suggests that the overlap issue is a problem that affects the quantitation process. In the next paragraph we are going to see how much does the overlap affect the computation of the ratio: in particular the case studied to evaluate the incidence of the problem shows an overlap between the first peak of the Heavy distribution and the seventh peak of the Light.
A general consideration about the results obtained by MaxQuant is that the quantification process is well performed. It quantified about the 96% of the peptides identified, and the average value is close to the expected one. In general, the only spot where our algorithm can make the difference, is in the standard deviation value, trying to decrease it. As we will see, it will be done. Moreover, it would be interesting to show an increase in the accuracy of the results for the peptide without the Lysine, a limited subgroup of peptides where, as just seen, MaxQuant doesn't perform any kind of correction for the overlap issue.

Incidence of the Overlap Problem

To perform this kind of analysis, three steps are necessary (fig. 5). The first part, as explained in the overview of this work, is the digestion of the whole proteome of an organism,. The organism used for our simulation is mouse (*Mus, Musculus*).

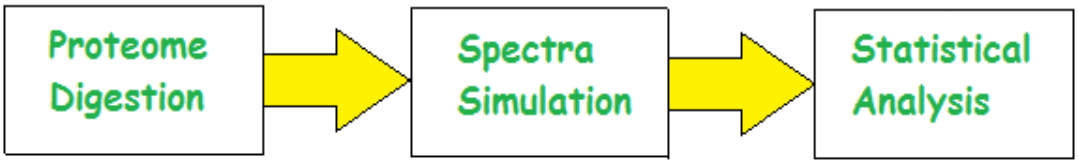


Fig.5: The operative workflow of the analysis of the incidence of the overlap

As known, the digestion is performed using the trypsin as cutting enzyme. The tool used for this simulation is named 'Protein Digestion Simulator' [22], whose screenshot is shown in the figure 6.

Once the proteome has been digested, it is necessary to collect a relevant number of peptides, and then compute their isotopic distributions on the basis of their mass. The numbers of peptide considered is 64,447, and the tool used for the isotopic distribution is IPC (Isotopic Pattern Calculator), already mentioned in the previous chapters. To compute the incidence of the overlap issue, two simple approaches have been realized:

1. the first one is based on three different analysis, which are simple and fairly similar. They consider the incidence of the overlap in terms of isotopic distributions, measuring the entity of the overlap;
2. the second one is based on the analysis of the 7^{th} peak, which is involved in the overlap: in particular, it has been considered the plot of the mass versus the relative intensity, and then it has been computed the number of peptides whose intensity is higher than some thresholds (higher is the mass, higher is the intensity of the peak).

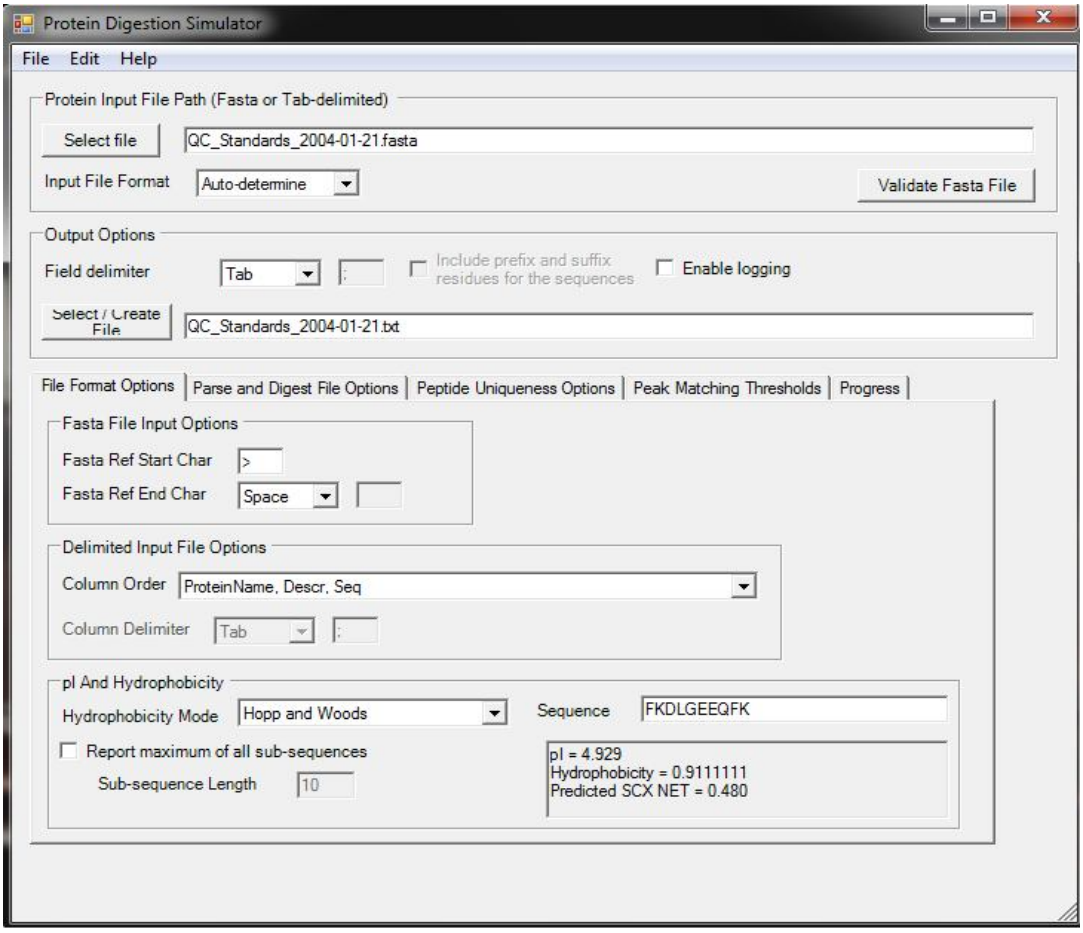


Fig. 6: a screenshot from Protein Digestion Simulator

First of all, it is important to understand how often the overlap due to the shift occurs in our set of peptides. In another way, we are trying to determine how many peptides have a seven Dalton long distribution. In that case, indeed, the problem of the overlap doesn't occur at all, because the distribution of the peptides isn't long enough to be superimposed. In the collection of 64447 peptides, 10451 are short enough to avoid the problem. This means that the percentage of non-overlapped peptides is:

`10451/64447 = 16.22%`

This percentage is quite low, and it allows us to glimpse that the overlap problem has an important impact on the computation of the Heavy to Light ratio.

The remaining 83.78% of the peptides, therefore, have an overlap problem. It is important to try to understand how it happens, to what extent and as widely. To do so, let's check out the three methods implemented. It is important to highlight that

the overlap considered in this part, is obtained with two distributions with the same intensity: we consider the Heavy to Light ratio equal to 1:1. This shrewdness allow us to check the incidence quite easily, as shown in the first method. Furthermore, the percentage reported are computed not on the total number of peptides (that is 64447), but only on those that overlaps (53996 peptides).

The first idea simply imposes a ratio between the maximum value of the distribution not overlapped (usually the first peak), and the maximum value of the overlapped distribution. It is easy to figure out that, if the value of the ratio is equal to one, it means that the two maximums are the same, and therefore there is no overlap at all.

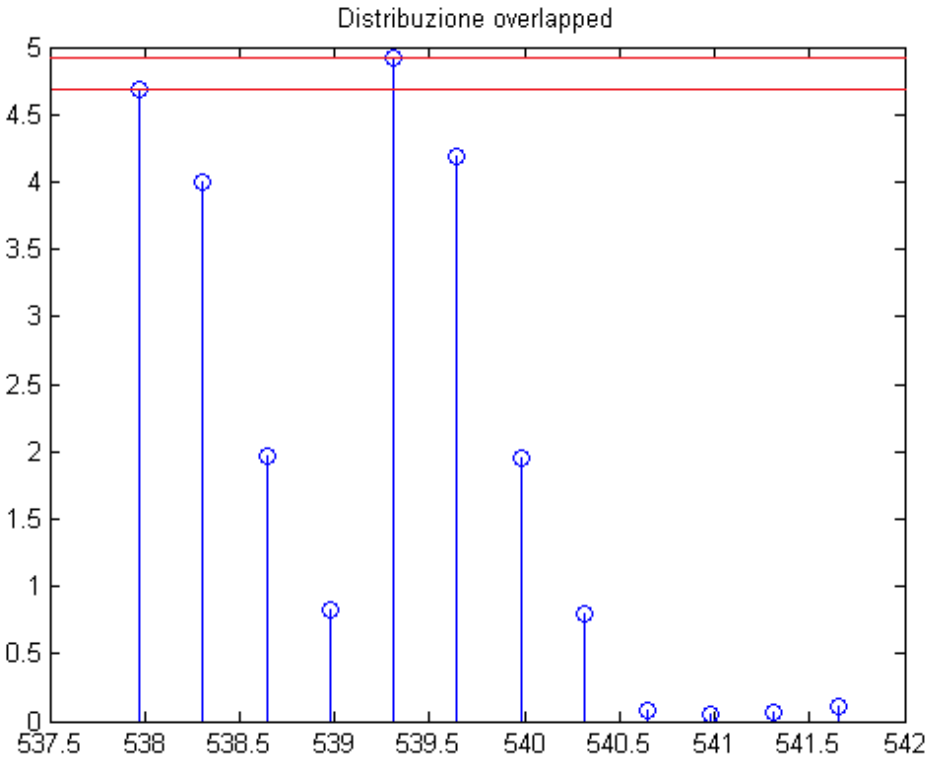


Fig.7: The first idea shown in a plot

Using this simple method it is possible to get some interesting data, that we can summarize in the next table. In particular, the ratios obtained have been split in four subgroups, according to their values. The first group has a value between 0.75 and 0.85, the second 0.85 – 0.95, the third one 0.95 – 0.97 and the last one 0.97 – 1.00. The percentage obtained are shown in the table:

0.75 < Ratio < 0.85	0,41%
0.85 < Ratio < 0.95	10,17%
0.95 < Ratio < 0.97	7,05%
0.97 < Ratio < 1.00	82,37%

Tab 3: the results of the first analysis

From this table it is possible to see that the most of the peptides have a ratio quite close to one, meaning that the overlap should not be so considerable.

The second method focuses particularly on the number of the peaks of the distribution which overlaps, regardless of their intensities. Therefore, we are trying to understand how many peaks interact each other independently from their heights. The results are summarized in the table:

1 < R < 3	48,74%
4 < R < 6	28,27%
7 < R < 9	13,89%
10 < R < 12	6,51%
13 < R < 15	2,49%
16 < R < 18	0,01%

Tab 4: the results of the second analysis

From this table emerges clearly that, in spite of what is emerged from the first method, the overlap is present and it involves many peptides. Indeed, almost half of the peptides have an overlap between one and three peaks, and more than the half have a number of overlapped peaks higher than three. It means that, even if the overlap is not very intensive, it is however quite widespread, introducing a considerable bias.

Finally, let's consider the last idea used to quantify the incidence of the overlap. This final analysis is quite similar to the first one, but instead relating the heights of the peaks, it relates the areas: in particular it has been computed the ratio between the area where the overlap occurs, and the area of the overall distribution. (fig.8)

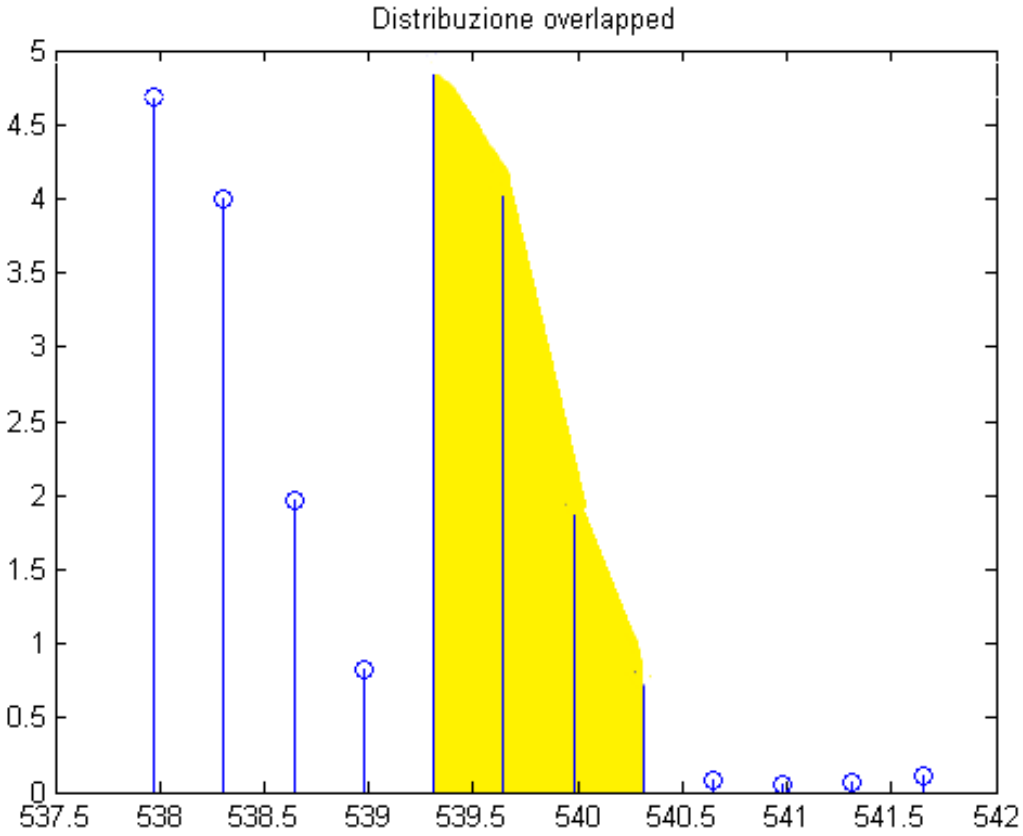


Fig.8: The areas from the third analysis

We may expect that if a little overlap occurs the final ratio is quite small, while if the overlap is considerable, the ratio is going to be bigger. In this third method the aim was to fuse the aspects of the two previous method: the intensity of the overlap highlighted in the first method and the numbers of peaks involved, as shown in the second method. Let's see the table of the results of the third idea:

0.00 < R < 0.20	19,00%
0.20 < R < 0.40	37,32%
0.40 < R < 0.60	38,55%
0.60 < R < 0.80	05,13%
0.80 < R < 1.00	00,00%

Tab 5: the results of the third analysis

In this table we can see that the mode is in the middle of the distribution of ratios: the 38.55% of the peptides have a ratio between 0.4 and 0.6. This method shows pretty well the final idea of the incidence of the overlap: the overlap imposes his presence quite widely in the distribution of the peptides, and may affect considerably the computation of the final ratio, even if the height of the peaks which overlaps isn't, on average, so intense.

To point up the intensity of the first peak of the overlap, the second part of the statistical analysis has been focused on the 7^{th} peak, which is the one that is overlapped with the 1^{st} peak of the labeled distribution. In this analysis, the aim was to understand how intensive is, on average, the 7^{th} peak, in order to understand how this affect the overall distribution: indeed, the trend of the intensity of this peak is a very interesting index about the problem of the overlap. Somehow this analysis tries to retrace the results obtained with the first method (about the ratio of the maximum of the distributions) but using some plotting function to visualize the role of the mass of the peptide in the value of the intensity of the 7^{th} peak. As it can be seen, higher is the mass of the peptide, higher is the intensity of the 7^{th} peak, therefore the problem of the overlap is more remarkable when the considered peptide has a big mass: it is possible to see again the importance of the digestion step, in order to get the right size in the sequence length. In the upper part of the figure 9, there is the approximation obtained with a 5^{th} grade equation, which best approximates the data; in the lower part there is the cluster of the real data: the range of the mass sweeps from about 400 Dalton up to 10000 Dalton. It is clearly visible that the trend increases with the mass with a function that seems to be like a sigmoid function. In the next table, tab. 5, there are some

numerical values which are extracted from this function, which show the numbers of peptides above the four stated threshold. These threshold are the 5%, 10%, 20% and 30% of the maximum value of the relative intensity.

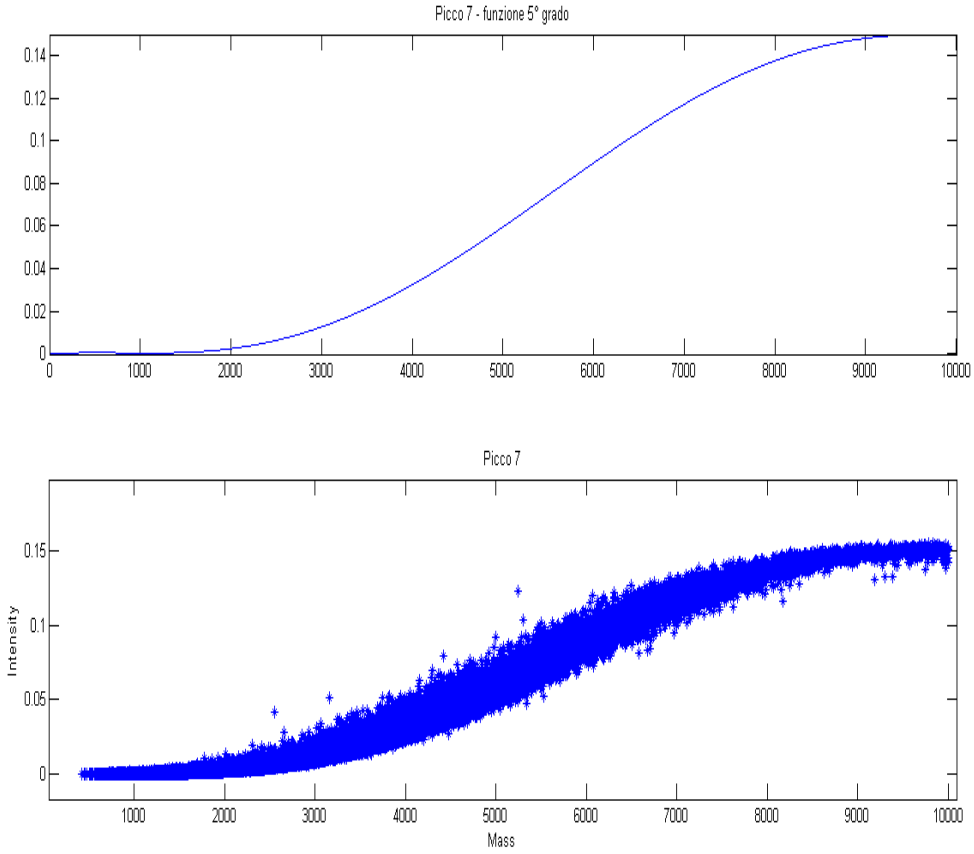


Fig.9: The trend of the 7th peak

If we consider a very small threshold, such as 5% of the maximum value, we get that almost 34% of the peptides exceed such threshold. Increasing the value of the threshold, obviously the number of peptides decrease, up to 16.50% of peptides for a threshold of the 30%. In the table 5 are shown the numerical values.

% Relative Intensity	Relative Intensity	Mass threshold 7° peak	# Peptides above threshold for 7° peak	% Peptides above threshold for 7° peak
0.0500 %	0.0075	2599.4	21562	39.93%
0.1000 %	0.0150	3143.8	16696	30.92%
0.2000 %	0.0299	3895.1	11700	21.67%
0.3000 %	0.0449	4487.2	8912	16.50%

Tab 6: the results of the 7th peak analysis

Finally, out of curiosity, it is reported an overview about the trend of all the peaks up to nine, simply to match the seventh peak with all the others (fig.10). We can easily see that there isn't a uniform trend between the peaks. Instead we can point out that the first peak of the isotopic distribution becomes lower with the mass, while the others increase. The last three peaks (included the seventh), instead, show an increasing trend, while the middle peaks (two to six) have an increasing trend up to a certain value, than a decreasing trend.

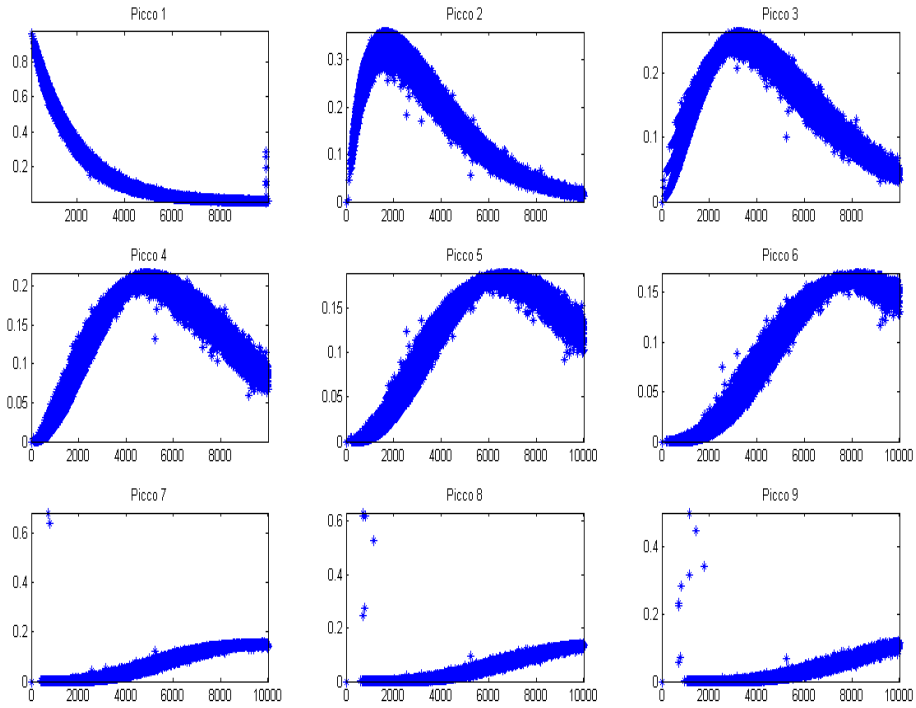


Fig.10: The trend of all the peaks up to 9

Results obtained through the proposed algorithm on the dimethyl datasets

In this paragraph the results obtained with our algorithm are shown, using the Dimethyl dataset. This section can be divided in four parts:

1. the analysis of the results from the peptides with at least one Lysine;
2. the analysis from those peptides without any Lysine;
3. a paragraph about the effect of tuning the parameters of the algorithm (threshold for the Pearson correlation and the threshold for the difference of the ratios); Finally, it is important to keep in mind that the expected value for this dataset is 0.5.
4. the results of the scoring on the dimethyl dataset.

Peptides with at least one Lysine

Let's see first how our first algorithm works with those peptides which have at least one Lysine in their sequence, and therefore don't have any issue due to the overlap problem. As already said before, the algorithm works taking as input the file '*evidence*' from the MaxQuant output, and it considers for the quantitation process only the desired peptides (those with at least one Lysine). The total number of peptides identified with at least one Lysine is equal to 8166.

The results obtained are shown in the table, together with the results obtained with MaxQuant (MQ column).
The first raw shows the value of the selected threshold for the difference between the three ratios. It is possible to see that when the difference is equal to one, the results are better in term of precision and recovered peptides.

Difference:	***MQ***	***0.5***	***1***	***2***	***5***	***None***
Average	0.5357	0.4678	0.5145	0.5594	0.6221	1.7416
Std	1.2063	0.3564	0.3941	0.4445	0.5791	31.9524
Quant	7869	4066	4744	4996	5129	6731

Tab.7: Results for those peptides with at least one Lysine

The threshold used for the Pearson correlation along the m/z axis is fixed and it is equal to 0.8 and, as we will see in the next paragraph (*'tuning the parameters: the best solution'*), this is the best value to get excellent results, both in the accuracy and in the numbers of quantified peptides. The second parameter, instead, is the threshold of the difference of the three ratios obtained from each part of the elution area (see the previous chapter for details). This parameter varies, as shown in the head of each column. Obviously, the smaller is the threshold, the better are the results obtained in terms of standard deviation, but smaller is even the number of quantified peptides. It is important to find out the right trade-off, tuning the parameter, in order to favor the quality of the quantification or the number of quantified peptides.

It is possible to see that, for the couple of parameters 0.8 (Pearson correlation) and 1 (difference of ratios), the standard deviation is highly reduced, but the number of peptides is significantly decreased. Therefore, we may state that the quality of the quantitation process is highly increased, at the expense of the number of quantified peptides. Let's see which are the results for the peptides with any Lysine amino acid.

Peptides without Lysine

For this category of peptides, the total number of triplet is equal to 1576. In the table, which is similar to the previous one, there are shown the results obtained.

Difference	***MQ***	***0.5***	***1***	***2***	***5***	***None***
Average	0.5503	0.4860	0.5339	0.5889	0.6580	1.0958
Std	1.4493	0.2059	0.2476	0.3326	0.4928	8.7881
Quant	1530	808	933	1002	1043	1488

Tab.8: Results for those peptides with any Lysine

As the previous one, the results are much better in terms of standard deviation, decreasing the value almost of an order of magnitude. As before, the price is in the

number of peptide quantified: from 97% to about 70%. We can say that this algorithm allows a great enhancement in the quality of the quantitation process, but it will lose a portion of peptides. Finally, it is possible to see that the Yoon's method works well, having a good precision (little standard deviation) and high accuracy (average value close to 0.5).
Let's see in the next paragraph how does the result change varying the values of the two key thresholds present in the algorithm.

Tuning the parameters: looking for the best solution

In this algorithm there are two thresholds, whose values are important to determine the accuracy of the quantitation process, and to determine the number of quantified peptides. As seen in the results, the accuracy and the number of quantified peptides are somehow in an opposite position: to have a good precision and a low standard deviation, some peptides have to be lost. In this paragraph it is shown how the number of the peptides recovered and the accuracy are affected by these two thresholds.
In the next two figures, there are shown two three-dimensional surfaces (sparsely sampled) where the axes of the plane are the thresholds of the algorithm: one for the Pearson correlation and one for the difference between the three sections of the elution area. The first one is the trend of the averaged ratio (therefore we are basing our discussion on the accuracy of the quantitation: how close to the expected value is the averaged value). It is already visible that the difference threshold (that one about the difference of the three ratios) is much more important to determine the precision of the result, while the Pearson correlation doesn't affect so much the precision of the quantitation process: therefore, if our aim is to achieve a precise quantification, we should tune the difference threshold, trying to reduce it as much as possible.

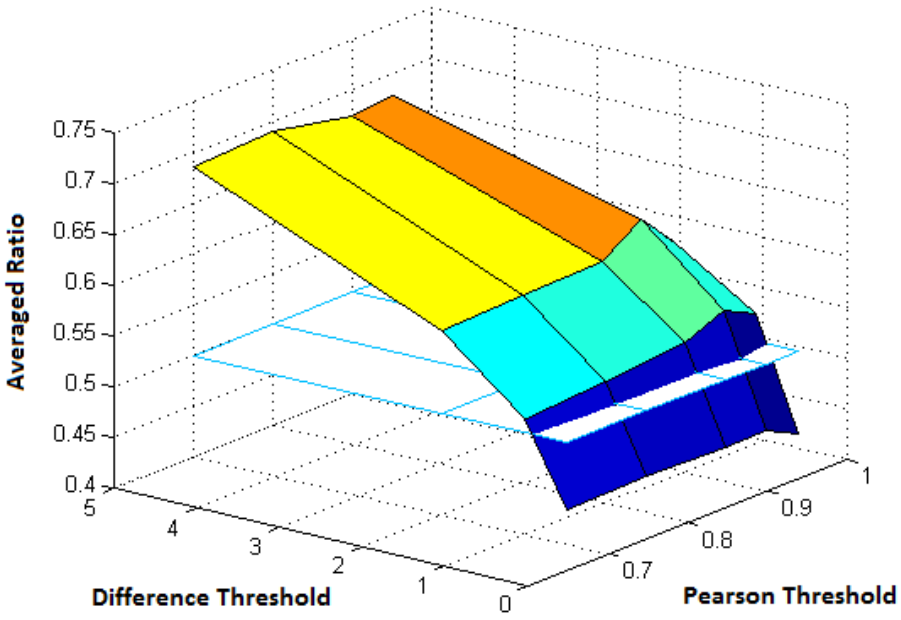
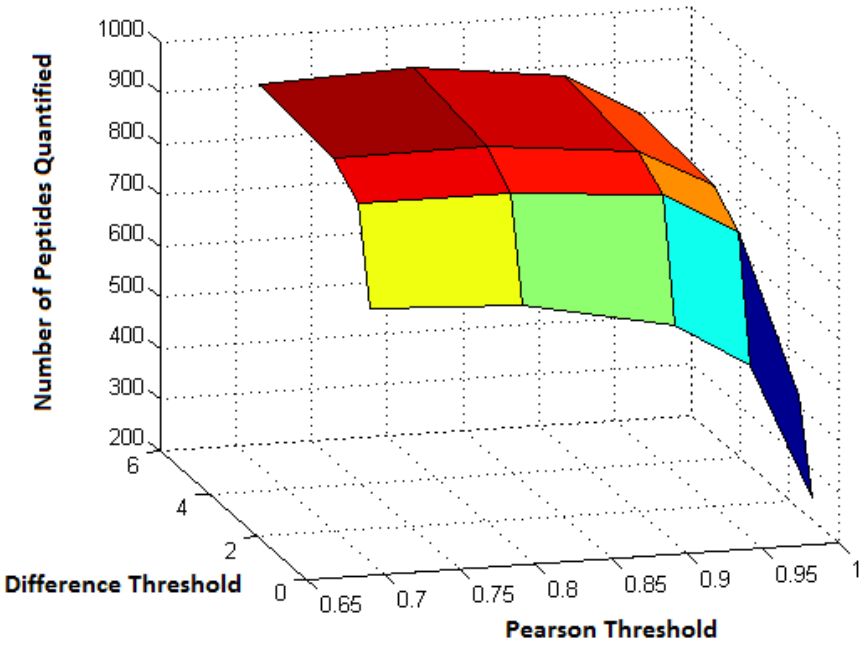


Fig.11: Averaged Ratio in the threshold domain



Tab.12: Numbers of peptides in the threshold domain

In the other figure (fig.12), it is possible to see the trend of the number of quantified peptides, varying the values of the thresholds. In this case, both the Pearson threshold and the Difference Threshold contribute at the final result: but the Pearson threshold has a much more important role in the number of quantified peptides. Hence, to increase the number of identified peptide, it is important to keep a low value of the Pearson threshold.

In the next two figures, the same concepts are shown in a bi-dimensional case, splitting each value of the difference threshold, and plotting on the x-axis the Pearson threshold.

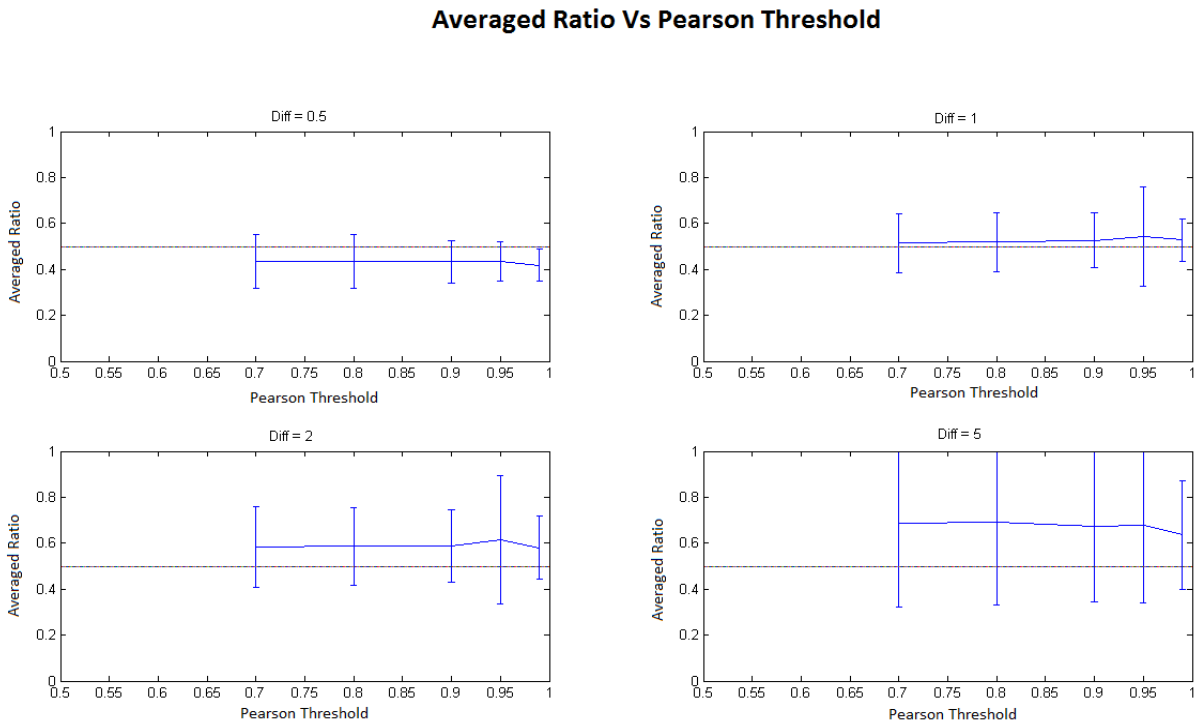


Fig.13: Average Ratio for each couple of threshold

It is possible to see that the precision and the accuracy of the quantification aren't related so much with the Pearson threshold, and the values obtained in each subplot of the first figure are quite constant. The precision is consistently dependent on the difference threshold rather than the Pearson threshold. Vice versa, the number of quantified peptides is related much more with the Pearson

coefficient, whose enhancement causes a sharp decrease after 0.9 values. On the other side, the difference threshold doesn't affect so much the number of quantified peptides, but it is anyway notable.

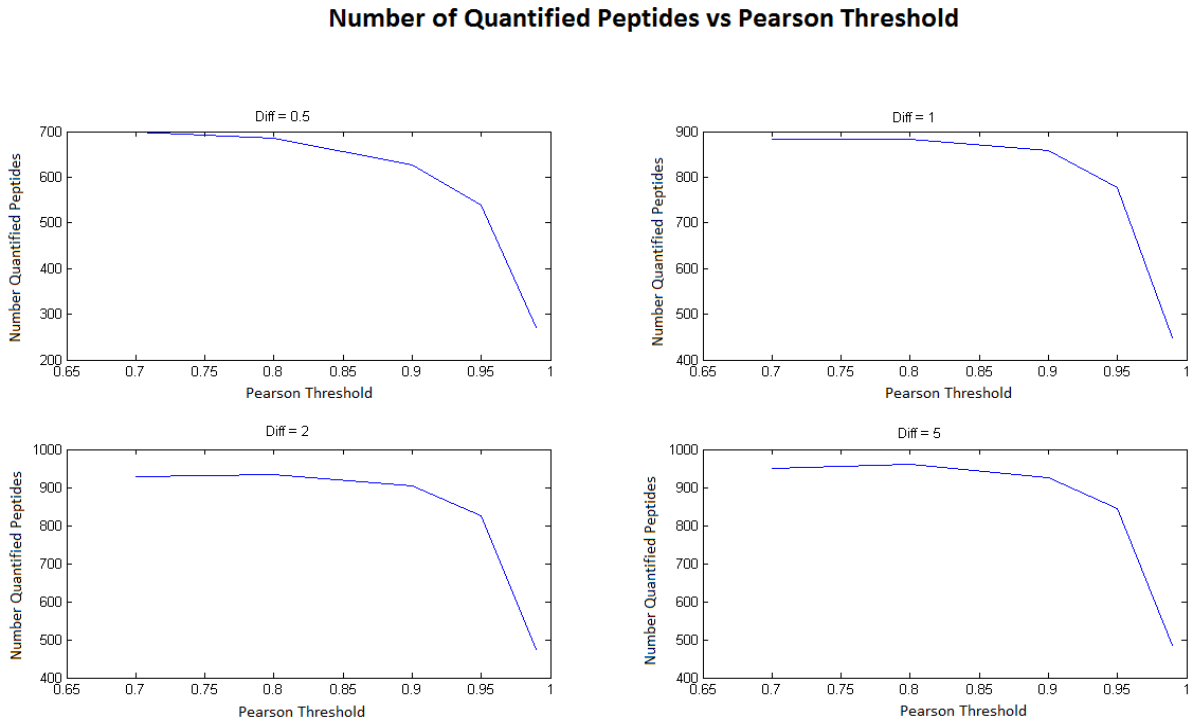


Fig.14: Numbers of quantified Peptides for each couple of threshold

Finally, the best strategy for a generic quantitation process, with a good accuracy and a consistent number of recovered peptides, is composed by such couple of values:

Pearson threshold = 0.8

Difference threshold = 1

An example: peptide 'AAASVMCHIEPDDGDDFVR'

In this part it is shown an example taken from a peptide of the Dimethyl dataset.

Sequence	‘AAASVMCHIEPDDGDDFVR’
m/z Light	711.65
m/z Medium	712.98
charge	3
Distance between peaks	0.33
Retention Time	103.76

Tab.9: Peptide’s characteristics

In the table 9, the main characteristics of the peptides are shown, such as the sequence (notice that there isn’t any Lysine, therefore the quantitation is performed using the Yoon’s method), the m/z position of the Light and of the Medium, the charge of the peptide, the distance between the peaks and finally the retention time. After the table, in the next figures, there are shown the 3D representation of the signal (fig.15): it is possible to see the elution area of the peptide: the fifth peak of the Light is clearly overlapped with the first peak of the Medium, at 712.98 Dalton. The same information is easily available from the bi-dimensional plot of the signal (fig.16), where the same peptide is looked along the m/z axis. Again, the overlap is clearly visible.

The ratio computed by the algorithm is very close to the expected value, which is 0.5. In particular, the three ratios are:

```
Ratio1 =       0.4876;
Ratio2 =       0.5145;
Ratio3 =       0.5233;
```

And the final ratio, computed as average of these three previous ratios, is:

```
FinalRatio = 0.5085
```

The threshold used for the Pearson selection of each scan is equal to 0.8, while the threshold used for the difference of the ratios is equal to 1. It’s interesting that the ratio computed for this peptide is really very close to the expected value, which is 0.5. As stated, due to the overlap, it has been used the Yoon’s method.

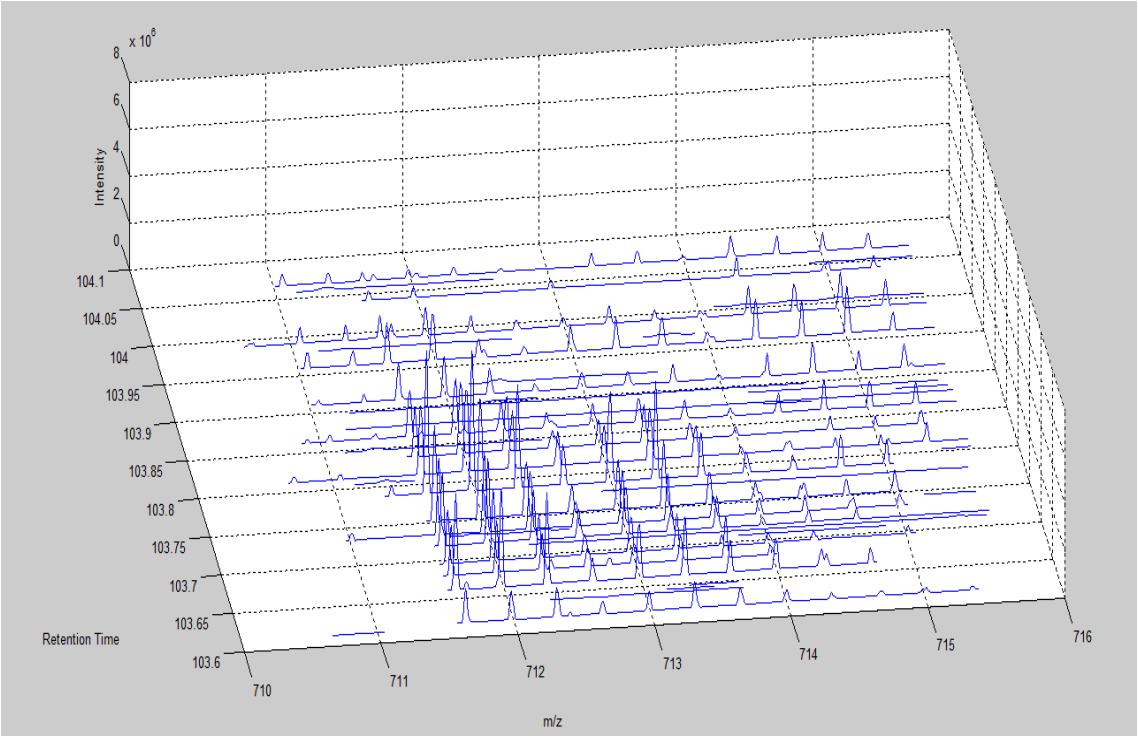


Fig.15: Three-dimensional representation of the signal

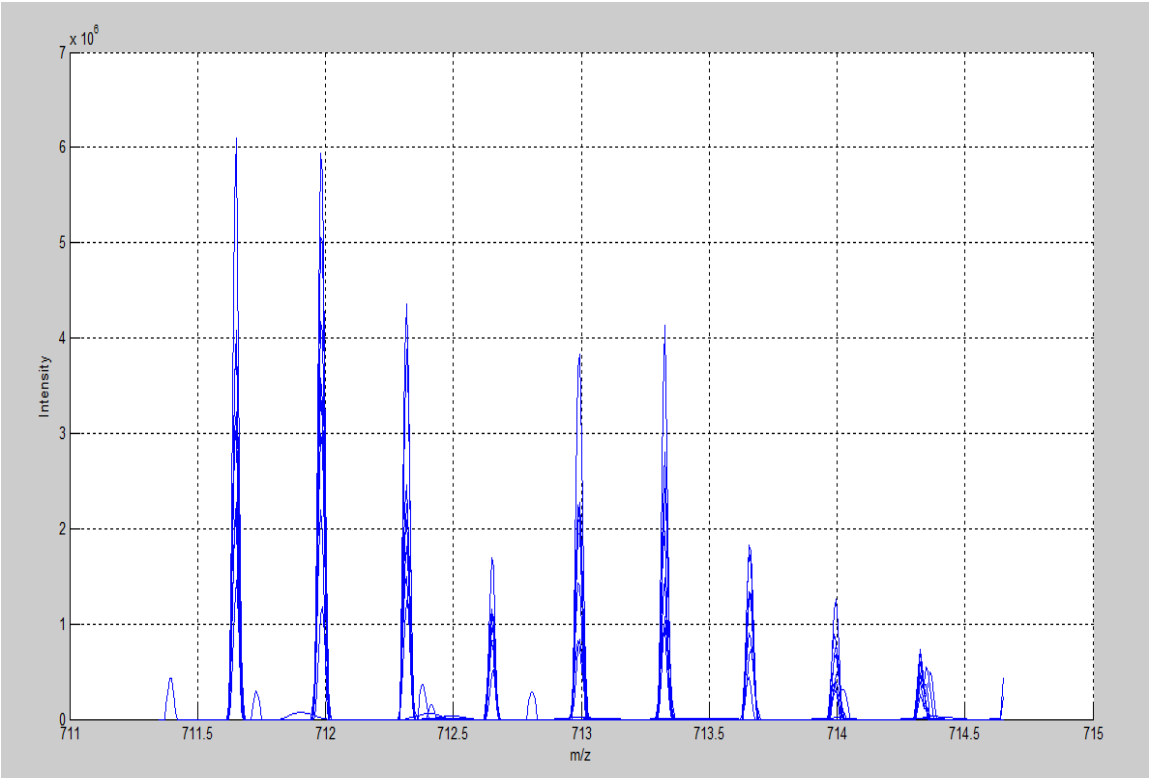


Fig.16: Bi-dimensional representation of the signal

Silac Dataset

Moreover, the algorithm has been tested on four different datasets, realized as well in the laboratory headed by the Professor Heck in Utrecht. Once again, the realization of such datasets has been realized on purpose, knowing the ratios within the triplet: two datasets have a proportion equal to 1:1:1, while the other two dataset have a 10:5:1 proportion. For these last datasets, the results have been obtained only for the first ratio Medium over Light, because of the trouble in obtaining the lower ratio 0.1. The enzyme used is again the Trypsin, cutting the sequence in presence of Lysine and Arginine. Unlike the dimethyl labeling, the Silac labeling has a different shift for the Lysine and the Arginine: the Lysine is shifted of 4 Dalton at each feature (as the previous dataset), while the shift of the Arginine is equal to 6 Dalton between the Light and the Medium, and 10 Dalton between the Light and the Heavy (therefore the Medium-Heavy difference is equal to 4 Dalton). In the next picture, it is shown a triplet with one Arginine in the sequence, and no Lysine. It is clearly visible the shift (keep in mind that the shift is always divided by the charge, that was in this case equal to 2). Another important issue that helps our algorithm, is about the absence of the retention time shift of the features, which is present in the Dimethyl labeling. Being well aligned, the elution area of the peptides are easily identifiable and quantifiable. The four datasets come from the same experiment: in particular two of them are the second fraction of this experiment, while the other two are the third fraction. The difference between dataset of the same fraction is in the ratios: as already said, the proportion are 1:1:1 and 10:5:1. The next table summarize the datasets.

DATASET	SILAC 02	SILAC 02	SILAC 03	SILAC 03
RATIOS	1 : 1 : 1	10 : 5 : 1	1 : 1 : 1	10 : 5 : 1

Tab.10: The datasets used for the second algorithm

Let's now see which is the performance of MaxQuant and Proteome Discoverer on these datasets.

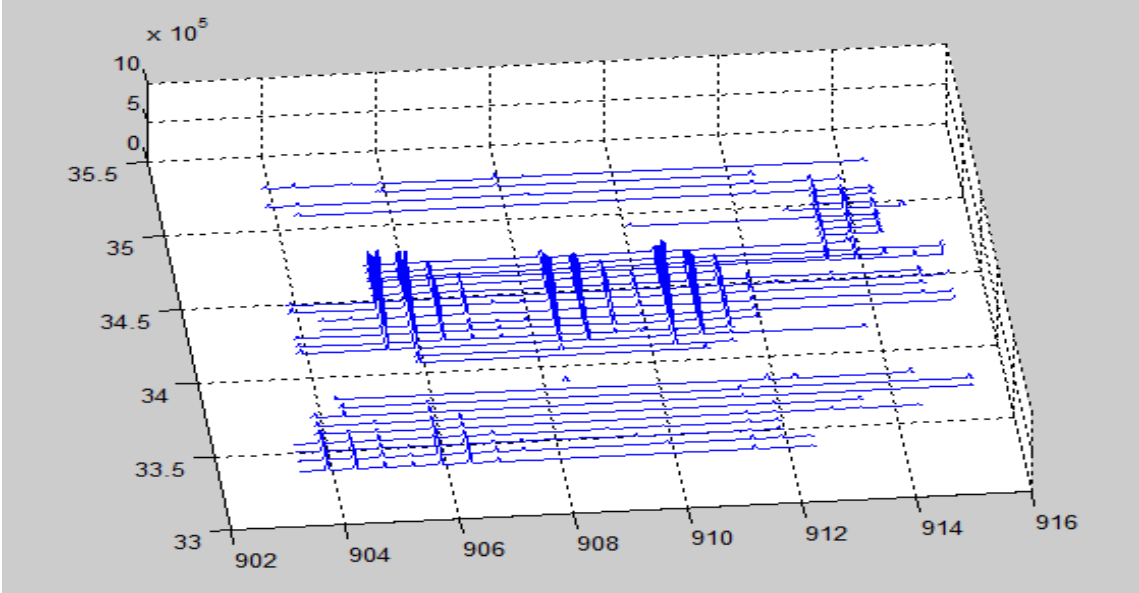


Fig.17: A triplet with an Arginine labeled

Comparing MaxQuant and Proteome Discoverer Results

The first difference between these two software is mainly based on the fact that MaxQuant is freeware, and realized at the Max Planck Institute of Berlin, while Proteome Discoverer is a software under license, realized by the ThermoScientific company. The comparison has been done under the same conditions: both use only one library for the identification process (MaxQuant uses Andromeda, while Proteome Discoverer, although might use a combination of several libraries, has been set to use only Mascot). In the next table there are shown all the results for both the software.

Dataset: **silac02 1:1:1**	MaxQuant		Proteome Discoverer	
Ratio:	ML	HL	ML	HL
Average:	0.746	0.847	0.8159	0.9945
Std:	0.711	0.630	0.9457	1.6714
Peptides Identified:	664	664	6903	6903
Peptides Quantified:	594	593	1074	1075

Tab.11: The dataset 1 results MQ vs. PD

Dataset: **silac03 1:1:1**	MaxQuant		Proteome Discoverer	
Ratio:	ML	HL	ML	HL
Average:	0.725	0.836	0.7494	0.8541
Std:	0.437	0.518	0.2123	0.3392
Peptides Identified:	8055	8055	27395	27395
Peptides Quantified:	7667	7666	15140	15518

Tab.12: The dataset 2 results MQ vs. PD

Dataset: **silac02 10:5:1**	MaxQuant	Proteome Discoverer
Ratio:	ML	ML
Average:	0.362	0.3482
Std:	0.157	0.0923
Peptides Identified:	646	4734
Peptides Quantified:	596	394

Tab.13: The dataset 3 results MQ vs. PD

Dataset: **silac03 10:5:1**	MaxQuant	Proteome Discoverer
Ratio:	ML	ML
Average:	0.408	0.3475
Std:	1.11	0.1498
Peptides Identified:	10494	23341
Peptides Quantified:	9943	7422

Tab.14: The dataset 4 results MQ vs. PD

The first important difference between MaxQuant and Proteome Discoverer is not about the quantification of the peptides, but it is about the identification process. The numbers of the peptides identified by Proteome Discoverer is much more bigger than the number of the peptides identified by MaxQuant. This difference in the identification process was already detected by Altelaar et al. [6] even at the level of the proteins. Aside from this identification issue, even the difference in the quantification are relevant. It is possible to divide the dataset on the basis of their fraction. In particular, in the fraction 2 (first and third tables) Proteome Discoverer

detected tenfold the numbers of peptides identified by MaxQuant, but the quantitation is very poor. In the first table, only the 15% of the identified peptides is quantified, and the results show a discrete standard deviation both for the Medium to Light ratio and for the Heavy to Light ratio. In the same fraction 2, with proportion 10:5:1, the quantification is even worse, because although identifying much more peptides than MaxQuant, the number of the quantified is incredibly smaller. Completely different is the fraction 3, in both cases. In the 1:1:1 case, the identified and quantified peptides are much more for Proteome Discoverer, and even the accuracy and the precision of the quantification is in favor of the commercial software, even if MaxQuant is able to perform fairly. The 10:5:1 dataset shows less peptides quantified, but an accuracy very high respect to MaxQuant, with a good precision for both of them.
In conclusion, Proteome Discoverer seems to be much better than MaxQuant, surely regarding the identification process; even for the quantitation process, generally Proteome Discoverer shows a better accuracy and a better precision than MaxQuant.
Let's see now if our algorithm is able to recover some peptide, or get a better standard deviation.

Results obtained through the proposed algorithm on the SILAC datasets

Our second algorithm has been tested on these four datasets. As for MaxQuant and Proteome Discoverer, in the dataset with a 1:1:1 ratios between the Light, Medium and Heavy labeling, we have computed the ratios Medium/Light and Heavy/Light, while for the 10:5:1 dataset, the only computed ratio is Medium/Light, because the feature of the Heavy distribution is close to the threshold of the noise. The results have been computed starting from the identification performed by the output file of MaxQuant (or Proteome Discoverer as well): in particular, MaxQuant provides also a number related with the Protein Group of that peptide, which means that we are able to compute the percentage of recovered proteins starting from the peptides.

The results are shown in the tables in the following pages. In particular, for each

table there are two columns: the first one shows the results obtained by MaxQuant, while the second one shows the results obtained using our algorithm. Let's finally see the table with the results obtained for the first dataset: the SILAC 02 1:1:1.

Dataset: **silac02 1:1:1**	MaxQuant		Our Algorithm	
Ratio:	ML	HL	ML	HL
Average:	0.746	0.847	0.752	0.817
Std:	0.711	0.630	0.328	0.430
Peptides Identified:	664	664	664	664
Peptides Quantified:	594	593	618	619
Protein Identified:	648	648	648	648
Protein Quantified:	587 90%	587	564 87% 96%	568

Tab.15: The dataset SILAC02 1:1:1 results

In the last row there are some interesting percentages. The first one shows the number of quantified protein up to the total number of identified proteins. It is possible to see that a relevant number of proteins are quantified. The second percentage shows the number of protein that our algorithm quantified, respect those quantified by MaxQuant. In this dataset it is interesting to see that the accuracy of the quantitation is better for our algorithm (both with and without the selection). In particular, the standard deviation is halved, and in the case without any selection the number of proteins lost compared to MaxQuant is only equal to 4%. In the next figure, it is shown the difference between MaxQuant and our algorithm, in a plot of the ratios versus the mass of the peptides. It is possible to see that our algorithm basically hasn't any outlier in its distribution of values, and all the ratios are close to the expected value. Even MaxQuant has a distribution close to the expected value (even more than our algorithm), but some peptides are clearly poorly quantified, leading to some outliers.

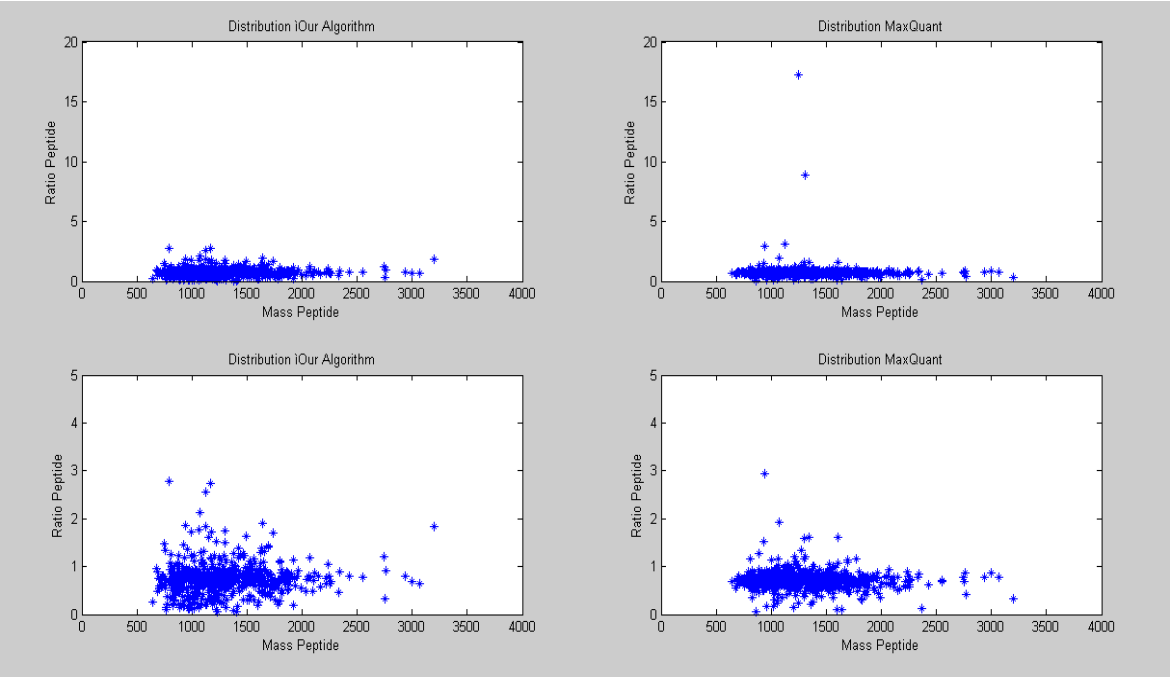


Fig.18: Comparison between MaxQuant and our algorithm

In the next tables are shown the results obtained with the other datasets: the SILAC 03 1:1:1, the SILAC 02 10:5:1 and the SILAC 03 10:5:1.

Dataset: **silac03 1:1:1**	MaxQuant		Our Algorithm	
Ratio:	ML	HL	ML	HL
Average:	0.725	0.836	0.782	0.804
Std:	0.437	0.518	0.416	0.523
Peptides Identified:	8055	8055	8055	8055
Peptides Quantified:	7667	7666	7521	7189
Protein Identified:	2796	2796	2796	2796
Protein Quantified:	2334 83%	2338	2233 79% 95%	2166

Tab.16: The dataset SILAC03 1:1:1 results

Dataset: **silac02 10:5:1**	MaxQuant	Our Algorithm
Ratio:	ML	ML
Average:	0.362	0.472
Std:	0.157	0.346
Peptides Identified:	646	646
Peptides Quantified:	596	574
Protein Identified:	622	622
Protein Quantified:	580 93%	532 85% 91%

Tab.17: The dataset SILAC02 10:5:1 results

Dataset: **silac03 10:5:1**	MaxQuant	Our Algorithm
Ratio:	ML	ML
Average:	0.408	0.531
Std:	1.11	0.471
Peptides Identified:	10494	10494
Peptides Quantified:	9943	9306
Protein Identified:	3613	3613
Protein Quantified:	2862 79%	2683 74% 93%

Tab.18: The dataset SILAC03 10:5:1 results

From these tables, it is possible to see that, except for the dataset SILAC 02 10:5:1, the standard deviation is always much better than that one of MaxQuant. After the selection of the peptides, the values of the standard deviation are always smaller than 0.3. Even the average value is close to the expected one (it has been seen that the real experimental ratios aren't equal to 1 and 0.5 but a little bit smaller: about 0.7 and 0.42).

Furthermore, as seen in the previous paragraph, the quantitation is much more reliable (smaller standard deviation) than MaxQuant, therefore our algorithm hardly can introduce any improvements. In the following tables the results obtained for the four datasets are reported.

Dataset: **silac02 1:1:1**	Proteome Discoverer		Our Algorithm	
Ratio:	ML	HL	ML	HL
Average:	0.8159	0.9945	0.6623	0.6790
Std:	0.9457	1.6714	0.4822	0.4943
Peptides Identified:	6903	6903	6903	6903
Peptides Quantified:	1074	1075	2764	2548

Tab.19: The dataset SILAC02 1:1:1 results

Dataset: **silac03 1:1:1**	Proteome Discoverer		Our Algorithm	
Ratio:	ML	HL	ML	HL
Average:	0.7494	0.8541	0.7178	0.7438
Std:	0.2123	0.3392	0.3813	0.4918
Peptides Identified:	27395	27395	27395	27395
Peptides Quantified:	15140	15518	14411	13641

Tab.20: The dataset SILAC02 1:1:1 results

Dataset: **silac02 10:5:1**	Proteome Discoverer	Our Algorithm
Ratio:	ML	ML
Average:	0.3482	0.3462
Std:	0.0923	0.2770
Peptides Identified:	4734	4734
Peptides Quantified:	394	1356

Tab.21: The dataset SILAC02 10:5:1 results

Dataset: **silac03 10:5:1**	Proteome Discoverer	Our Algorithm
Ratio:	ML	ML
Average:	0.3475	0.3873
Std:	0.1498	0.2629
Peptides Identified:	23341	23341
Peptides Quantified:	7422	9847

Tab.22: The dataset SILAC03 10:5:1 results

Proteome Discoverer has an excellence standard deviation, and our algorithm is able to perform better only in the first dataset. But the number of peptides quantified by our algorithm is significantly bigger in the first and in the third dataset, with comparable averaged values.

Some example: Peptide 'AAAAAAGEAR'

Let's now see an example of the whole quantitation algorithm for a well-quantified peptide, coming from one of the SILAC dataset, and whose sequence is 'AAAAAAGEAR'. In the next table are reported the data of the peptide: the m/z value of the Light, of the Medium and of the Heavy distribution, the charge and finally the distance between the peaks of the same isotopic distribution.

Sequence	'AAAAAAGEAR'
m/z Light	450.7303
m/z Medium	453.7303
m/z Heavy	455.7303
charge	2
Distance between peaks	0.5
Retention Time	37.5610

Tab.23: Peptide's characteristics

Being only one amino acid Arginine ‘R’ and no Lysine ‘K’, the shift between Light and Medium is equal to six Dalton divided by the charge, and the distance between the Light and the Heavy is equal to ten Dalton divided by the charge. Therefore the first peak of the Medium should be overlapped with the seventh of the Light (which, basically, doesn’t exist) and the first peak of the Heavy with the eleventh of the Light and the fifth of the Medium. The last information reported in the table is about the retention time of the fragmented ion (when the peptide has been identified).

In the figure 19 it is possible to see the whole triplet of the real signal. As it is possible to see, the signal-to-noise ratio (SNR) is very high, and there isn’t any kind of overlap with any other peptide. Furthermore, the theoretical distribution of the peptide has the fourth and the fifth peaks with a very low intensity, therefore there shouldn’t be any noisy overlap between the Light feature and the Medium, or between the Medium and the Heavy.

In the next figure (fig. 20), it is shown a combined plot composed by four different pictures, which are about the first peak of the Light distribution. In the first one there is the real signal, named *‘All the peak’*. Under the real signal there is the figure of the filtered signal: the real Gaussian is filtered with a low pass in order to be able to find the best fit with the Exponentially Modified Gaussian, which is shown in the picture to the upper right corner. Finally, in the fourth subplot, it is shown the signal used for the quantitation, obtained as described previously from the EMG and the filtered signal.

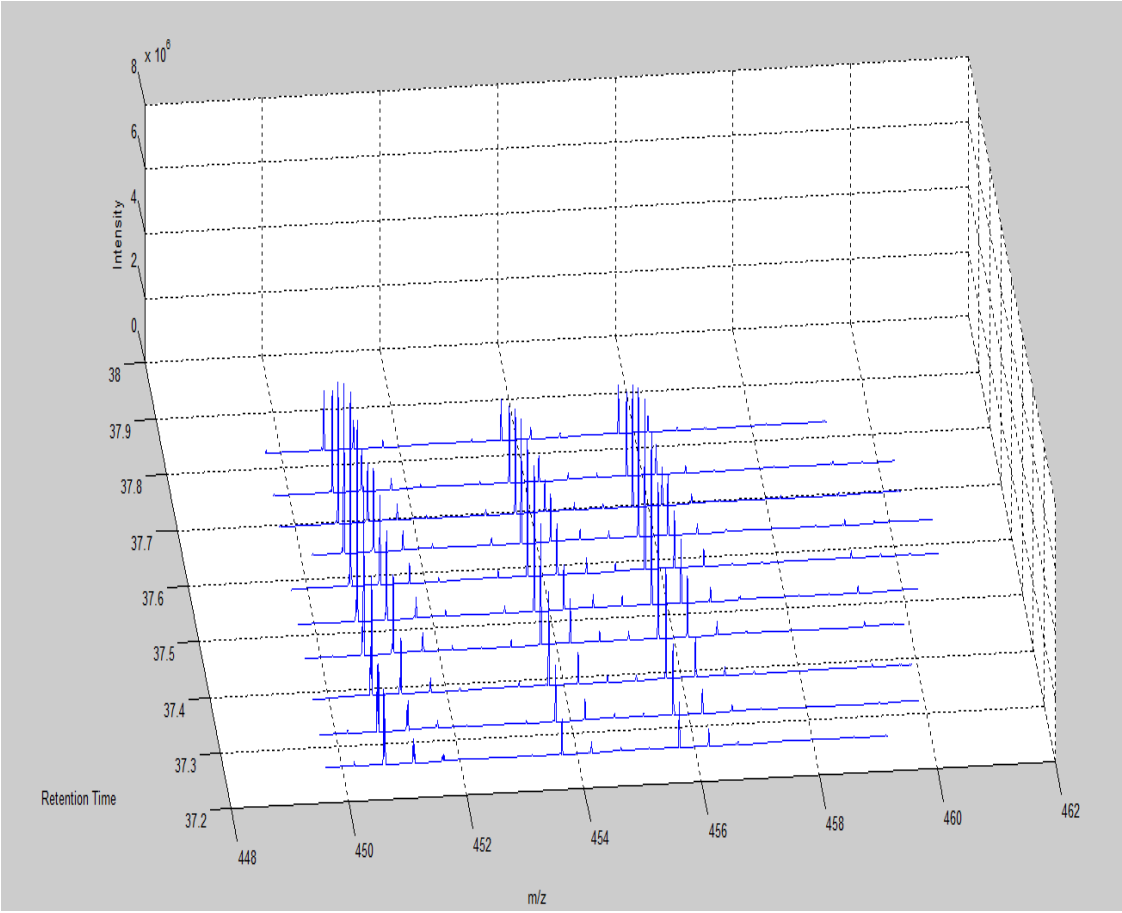


Fig.19: The whole 3D triplet

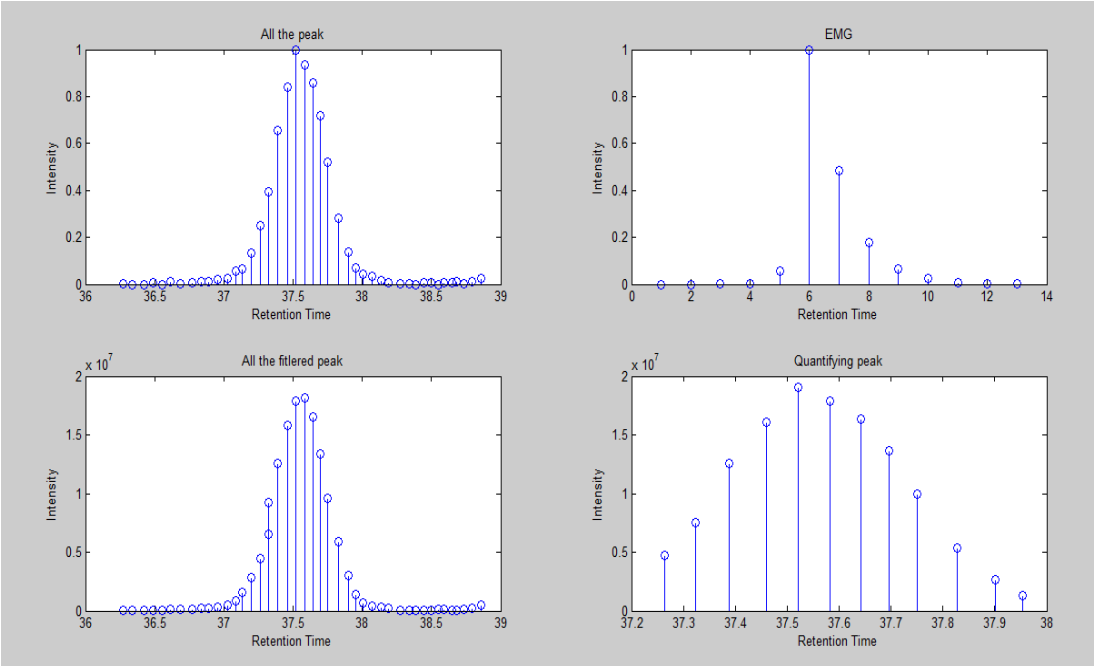


Fig.20: The subplot about the retention length problem

In the next three pictures (21, 22 and 23) are shown all the three peaks of the Light, Medium and Heavy, and their final division in five parts, differently colored.

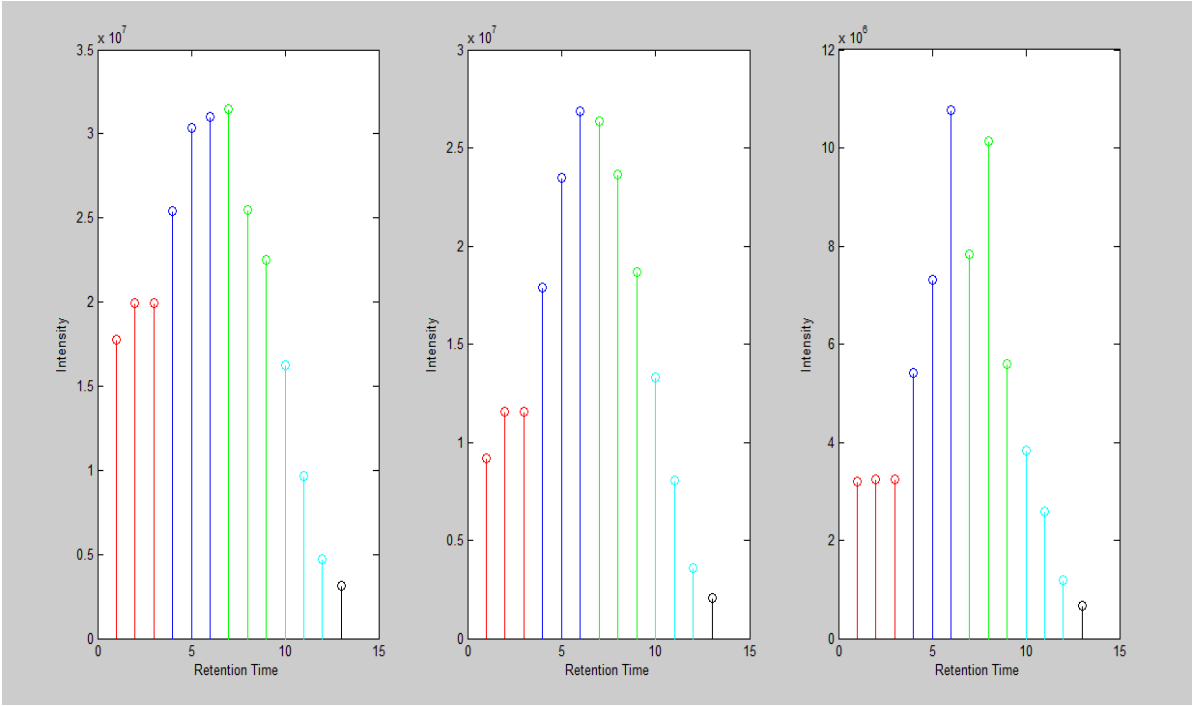


Fig.21: The Light peaks

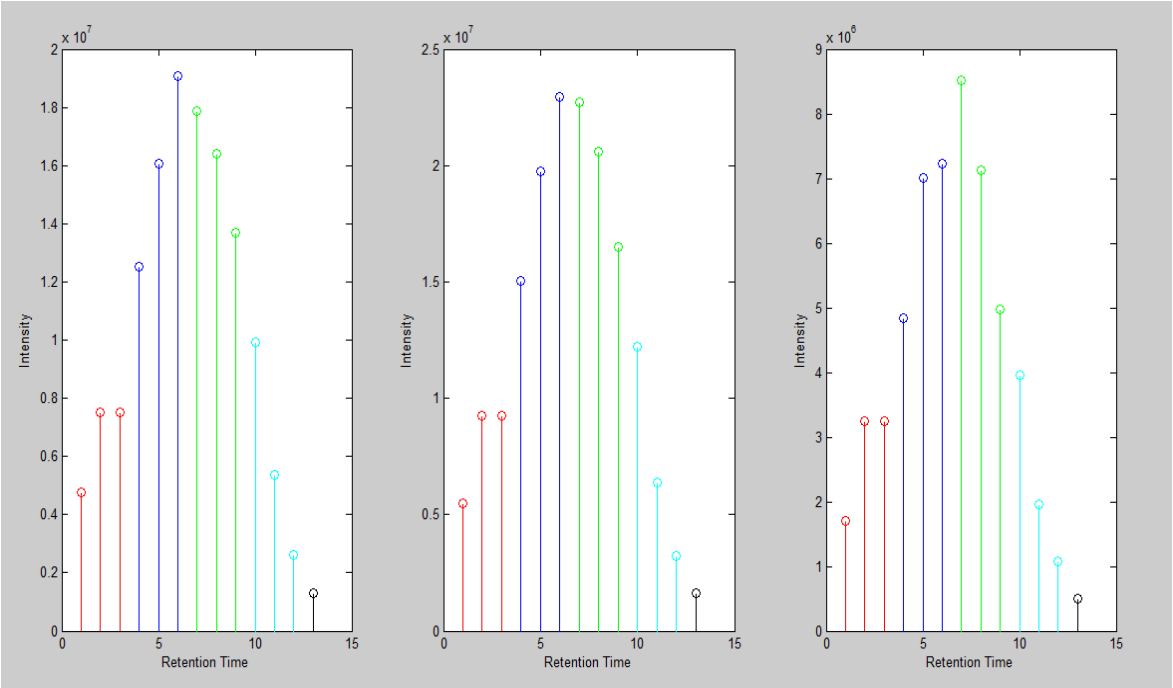


Fig.22: The Medium peaks

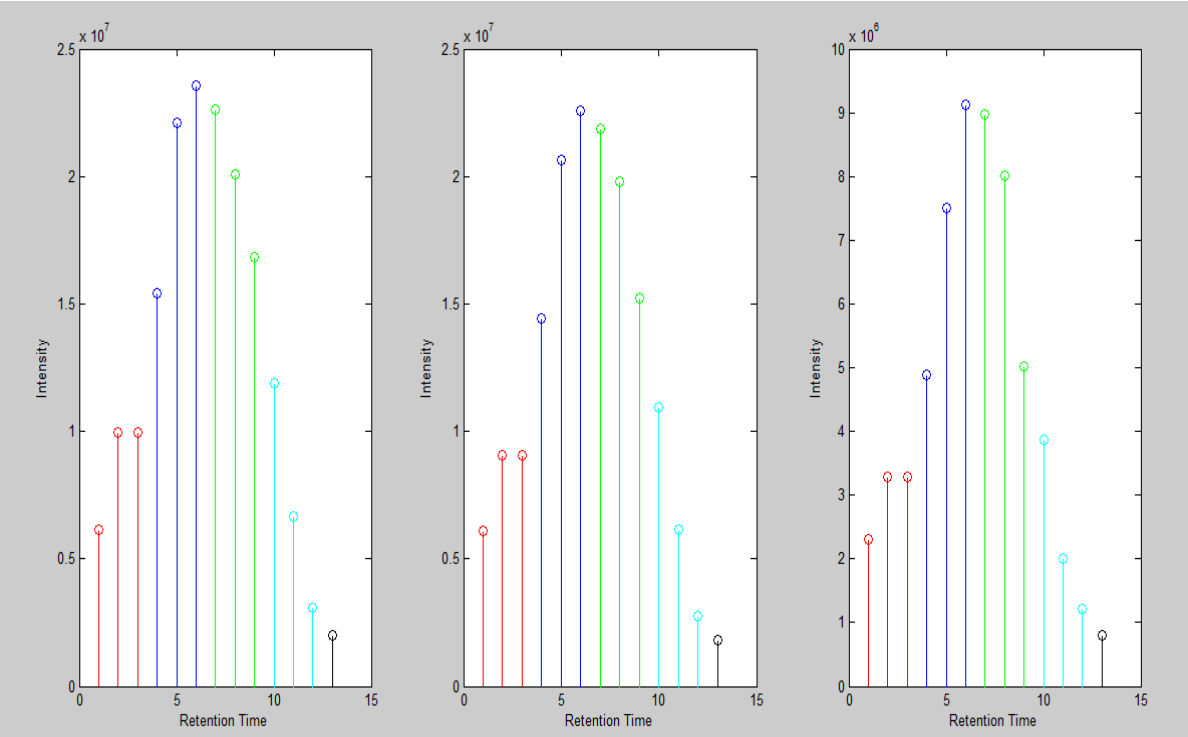


Fig.23: The Heavy peaks

Finally, there are reported the five values obtained from each part: both from the Medium / Light ratios and from the Heavy / Light ratios.

ML: 0.6437 0.6526 0.7362 0.7833 0.7931

HL: 0.7071 0.8113 0.8219 0.8375 0.9092

It is possible to see that such ratios are very close to each other, and therefore they belong to the same cluster. The final ratios is then obtained as average of the whole vector, and then we have as result:

```
ratioMLfinal =      0.7218;
ratioHLfinal =      0.8174;
```

Finally, we get the two scores obtained as combinations of the different contributions: the score Id, the length of the final vector, and the Pearson correlation in the m/z domain and in the Retention Time domain. The scoring values are:

```
scoreML =      0.8591;
scoreHL =      0.8589.
```

Some example: Peptide 'AAAVSSVVR'

The following examples are very important to prove the effectiveness of our algorithm, because they show the good results of our quantitation process, where MaxQuant fails. The first examples comes from the SILAC dataset, e its sequence is AAAVSSVVR. In the table are shown the most important characteristics

SEQUENCE	AAAVSSVVR
m/z Light	451.2587
m/z Medium	454. 2587
m/z Heavy	456. 2587
Retention Time	82.2390
Charge	2

Tab.24: Peptide's characteristics

The peptide may be hardly quantified due to the presence of other peptides, whose elution area is partially overlapped with that one of our peptide, especially in the heavy feature, as it is possible to see in the figure 24. For this reason, MaxQuant is not able to properly quantify the peptide, and the result of its quantification is null (expressed in the Matlab workspace with the notation NaN – which means Not A Number).

Our algorithm, dividing the area in five parts, is able to compute the final computation, obtaining a result very close to the expected one. The five ratios and the final one, both for the Medium to Light ratio and for the Heavy to Light ratio, are:

```
Ratio MaxQuant ML = NaN
Ratio Our Algorithm ML = 1.1451
(score 0.7807) ratios = [0.9218  0.9970  1.0494  1.1361  1.6213]

Ratio MaxQuant HL = NaN
Ratio Our Algorithm HL = 1.1997
(score 0.7172) ratios = [0.9578  1.1984  1.2054  1.2884  1.3487]
```

Finally, in the next figures, it is shown the division of the three peaks of the Light, Medium and Heavy distribution in five parts.

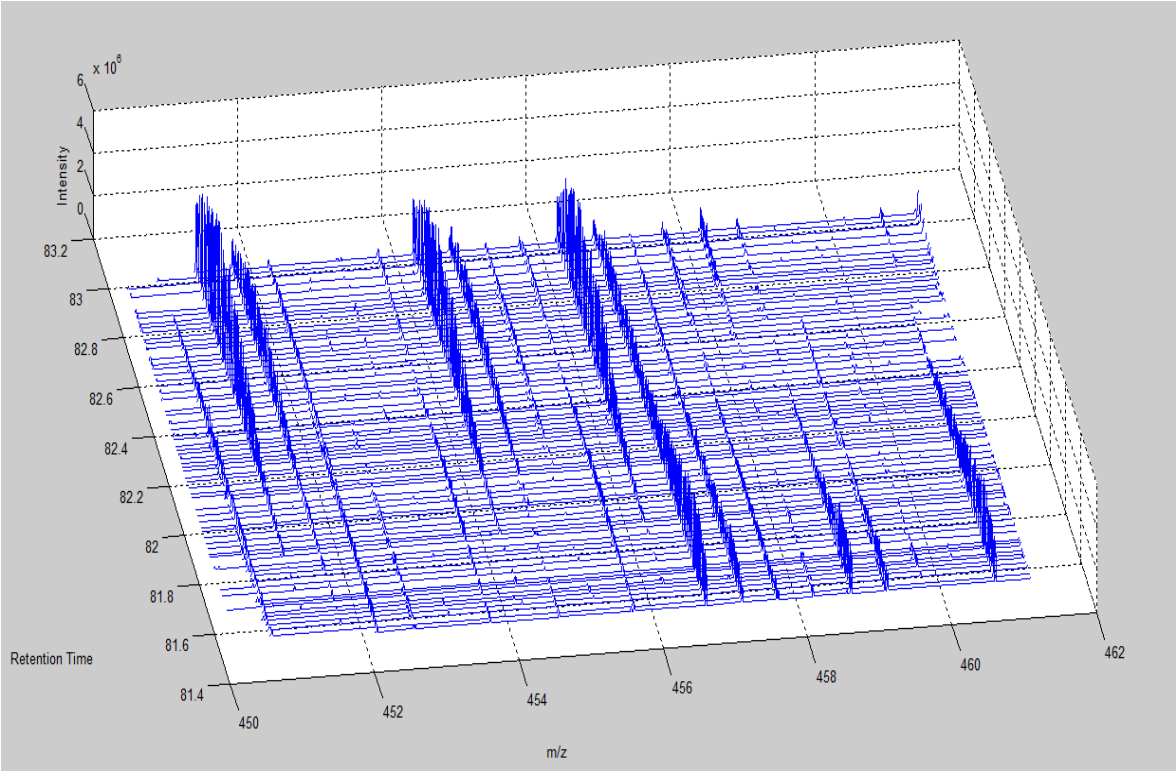


Fig.24: A crowed area all around our peptide

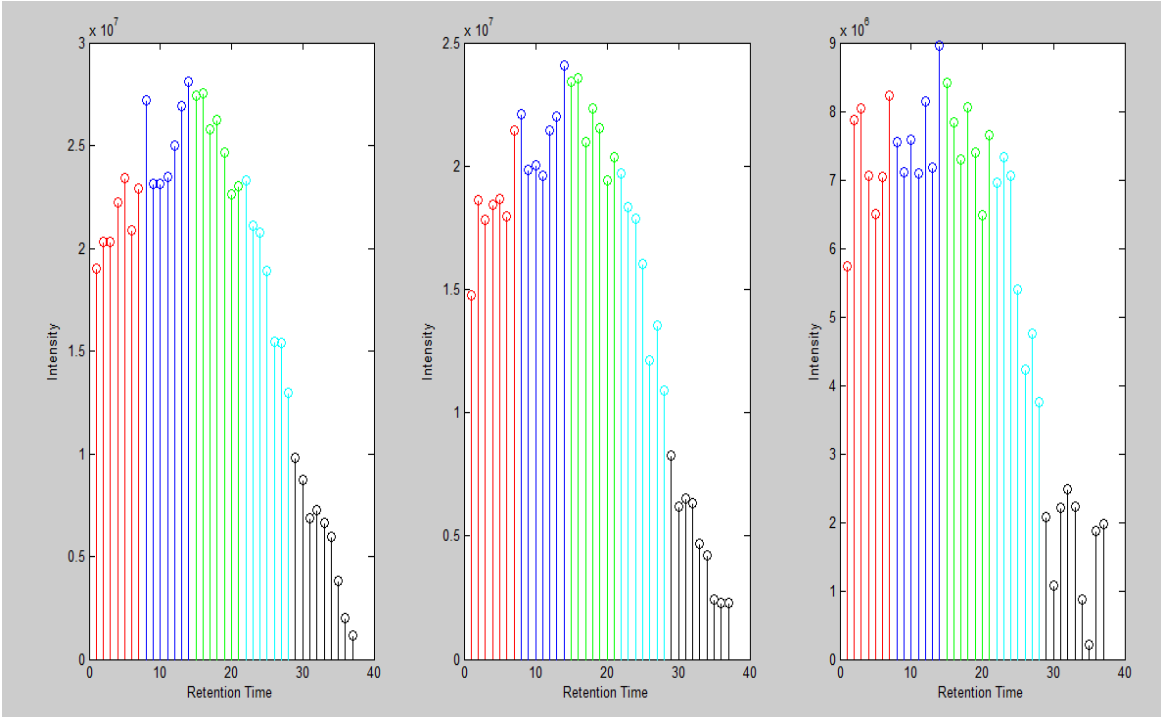


Fig.25: The division of the Light Peaks

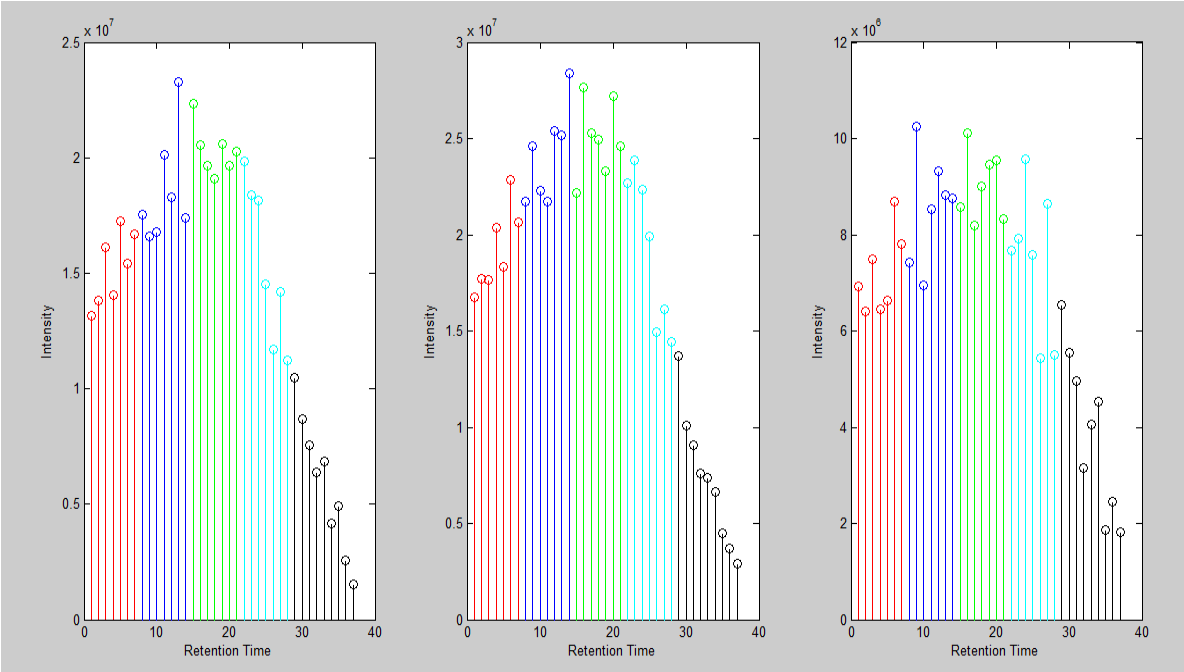


Fig.26: The division of the Medium Peaks

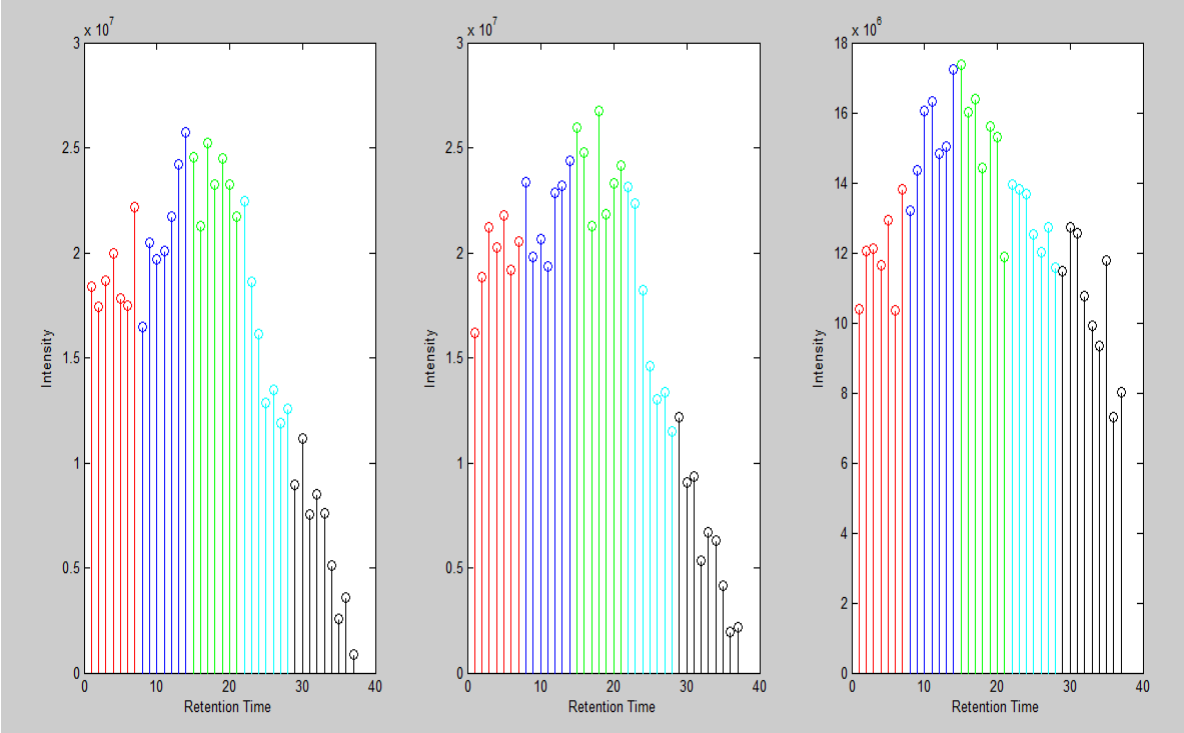


Fig.27: The division of the Heavy Peaks

Some example: Peptide ‘MISGERK’

Another peptide example, picked up from the SILAC dataset. In the table its characteristics.

SEQUENCE	MISGERK
m/z Light	431.7262
m/z Medium	436. 7262
m/z Heavy	440. 7262
Retention Time	64.5400
Charge	2

Tab.25: Peptide’s characteristics

In this case, the peptide MISGERK is localized in a very crowed region, and MaxQuant is not able to properly quantify it. Our algorithm, on the contrary, doesn’t fail.

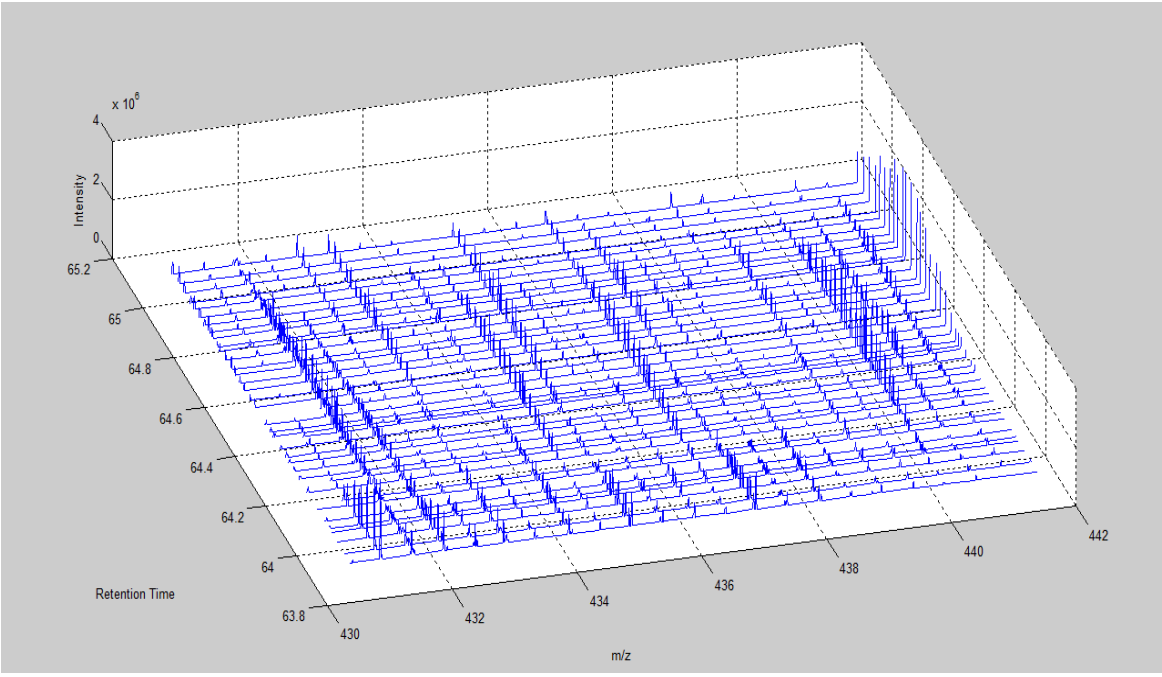


Fig.28: Crowed signal with our peptide

Let's see how MaxQuant quantify, and how our algorithm perform. It is possible to see that our algorithm provides the best ratio even for this crowed peptide.

```
Ratio MaxQuant ML = 0.0465
Ratio OurAlgorithm ML = 0.7605
```
(score 0.6519) ratios = [0.3900 0.6253 0.8451 0.8720 1.0702]

```
Ratio MaxQuant HL = 0.1379
Ratio OurAlgorithm HL = 1.5600
```
(score 0.4790) ratios = [0.2655 0.5061 0.5936 0.6048 0.8301]

MaxQuant underestimate significantly the final ratio. Our algorithm, instead, is able to get a value really close to the expected one.

Some example: Peptide 'SIFDIFR'

This examples, picked from the SILAC dataset, shows an important characteristic of our algorithm, and its reliability in the quantification process. The data of the peptide are shown in the following table.

SEQUENCE	SIFDIFR
m/z Light	470.2504
m/z Medium	473. 2504
m/z Heavy	475. 2504
Retention Time	156.590
Charge	2

Tab.26: Peptide's characteristics

In this case, the peptide SIFDIFR has an important overlap at 473.25, the spot of the medium feature, as shown in the picture of the signal, and this obviously affected the ratio computation.

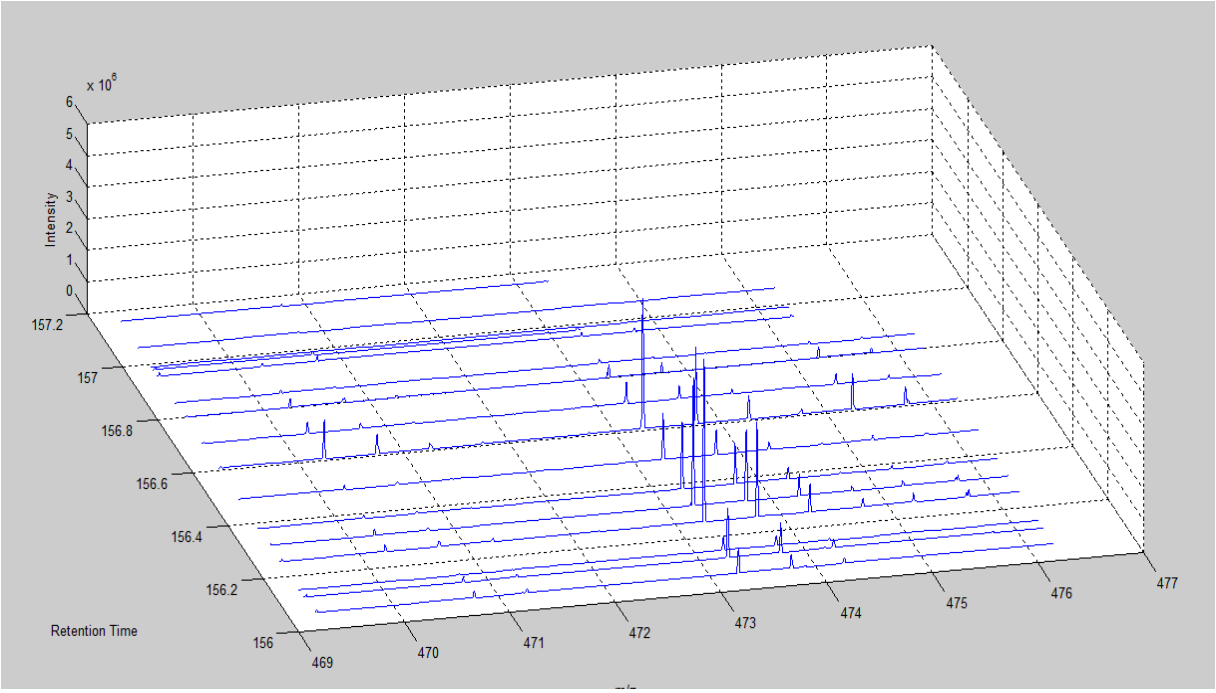


Fig.29: Overlapped feature at 473 Dalton

Such overlap is a problem, because it alters significantly the ratio Medium to Light, overestimating the value. As shown in the following, MaxQuant reports a value which is almost equal to three, while the expected value should be close to one.

```
Ratio MaxQuant ML = 2.9439
Ratio OurAlgorithm ML = NaN
(score 0.4493)ratios = [0       0    6.1870    7.2187    8.4383]

Ratio MaxQuant HL = 0.8601
Ratio OurAlgorithm HL = 1.0080
(score 0.7237)  ratios = [0    0.5875    0.9843    1.0438    1.4165]
```

Our algorithm detects the problem, because the difference between the ratios is too high, and therefore there should be a mistake in the computation. The strategy adopted in such case is not to quantify, not being able to provide a reliable value. For the other ratio, Heavy to Light, there aren't any problems, because the overlap is spread only over the medium elution area.

Scoring the quantification

The peptides of the first dataset (SILAC 03, 1:1:1) have been plotted in the three dimensional space composed by the Pearson coefficient of the peptides, the Identification Score and by the number of ratios used for the final quantification; these are the three elements which are used to compute the score of each peptide (see the previous chapter for details). They have been classified in two groups:

1. the first group is composed by those peptides whose ratios is in the range 0.5 – 1.5. The second group of peptides is composed by all the other peptides, whose ratio obviously isn't in that range.

Once the peptides have been classified (fig.30), we have performed a linear classification to separate the two different groups, through two linear discriminant analysis, weights for the three parameters were determined.

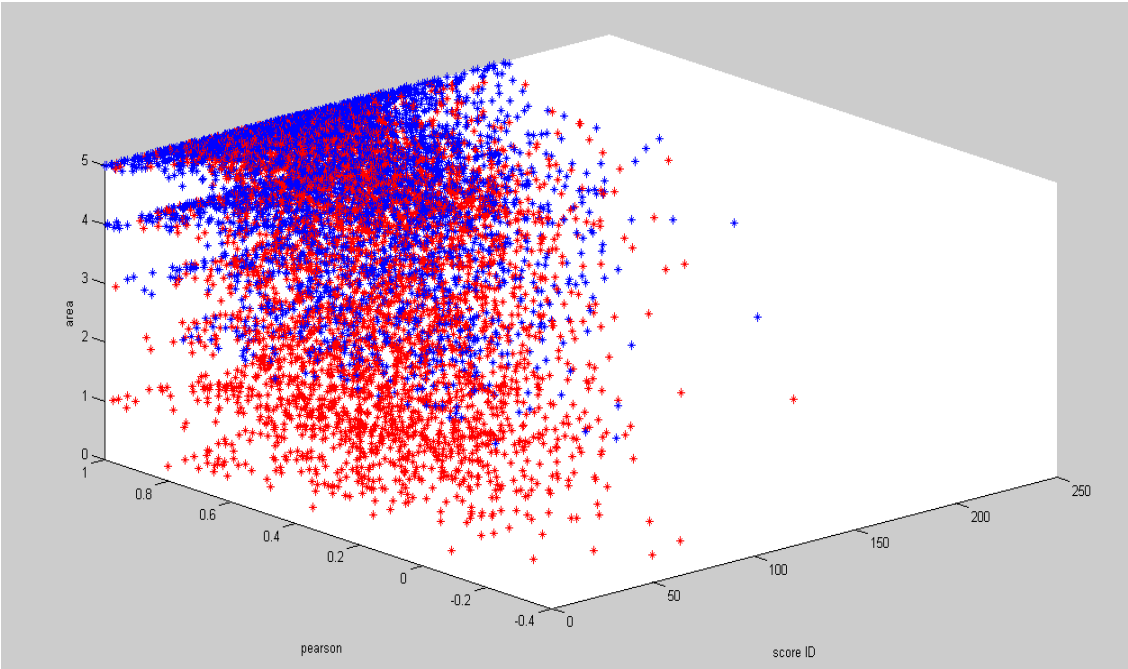


Fig.30: The three dimensional space and the classified peptides: the blues have a ratios between 0.5 and 1.5.

The result of the separation is shown in the next figure (fig. 31).

It is possible to see that the linear separation privileges those peptides whose

position is in an upper corner of the space. It means that those peptides with high values of the components show a good ratio (where good means a ratio within the specified range). This is basically what we expected. Exploiting this information, a peptide may be provided with a score relative to quantification, analogously to what happens for identification. In this way, we have assigned a score to each peptide of the datasets.

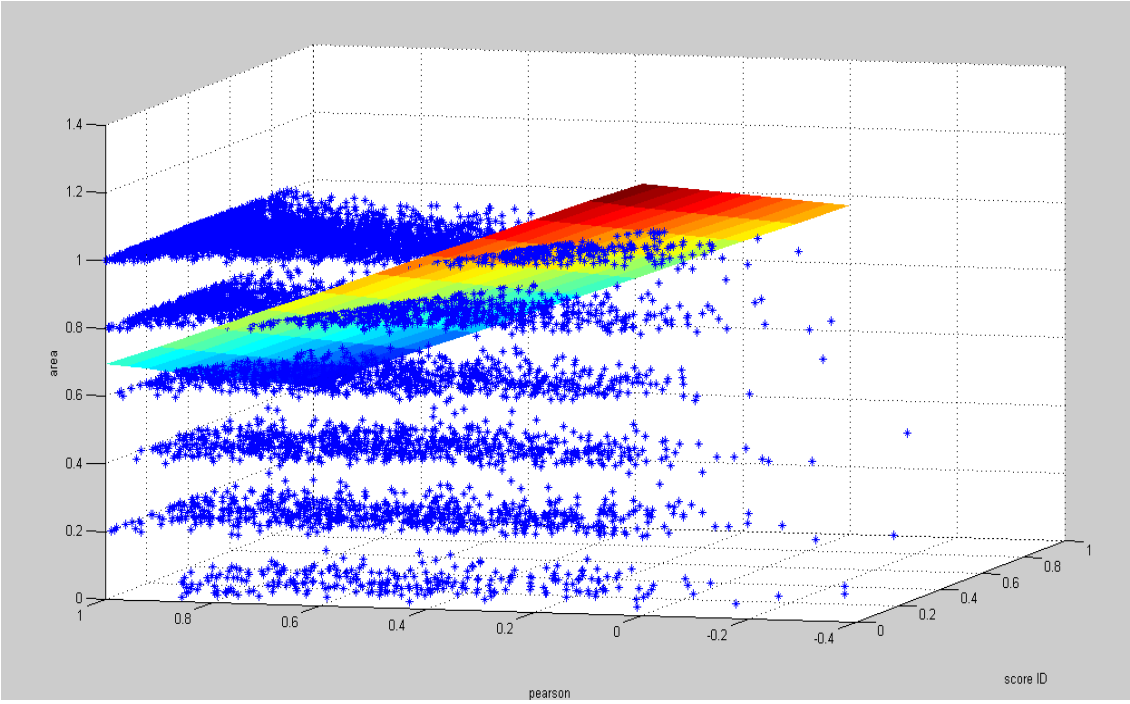


Fig.31: The linear separation

As it is possible to see from the next pictures (fig 32 to 34), as the threshold of the score increases, as the standard deviation of the results of the quantification decreases. Similarly, the number of quantified peptide decreases. Therefore, it is possible to select peptides according the quality of quantification.

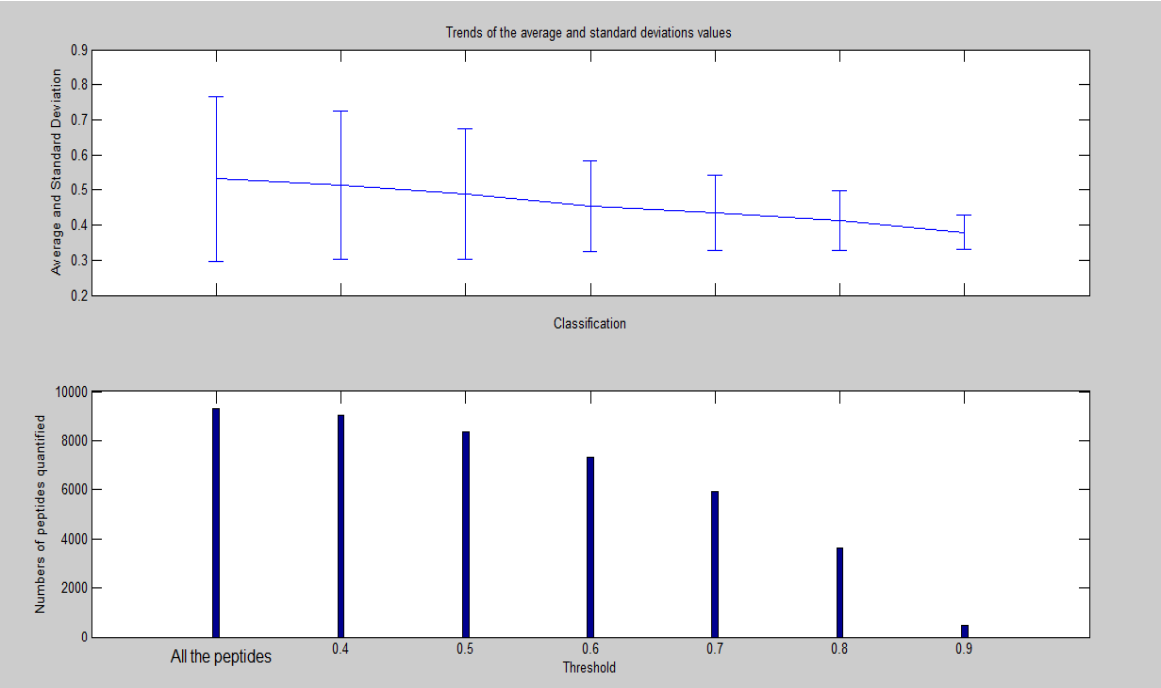


Fig.32 Classification and thresholding of the Medium/Light ratio in the Dataset SILAC02 10:5:1

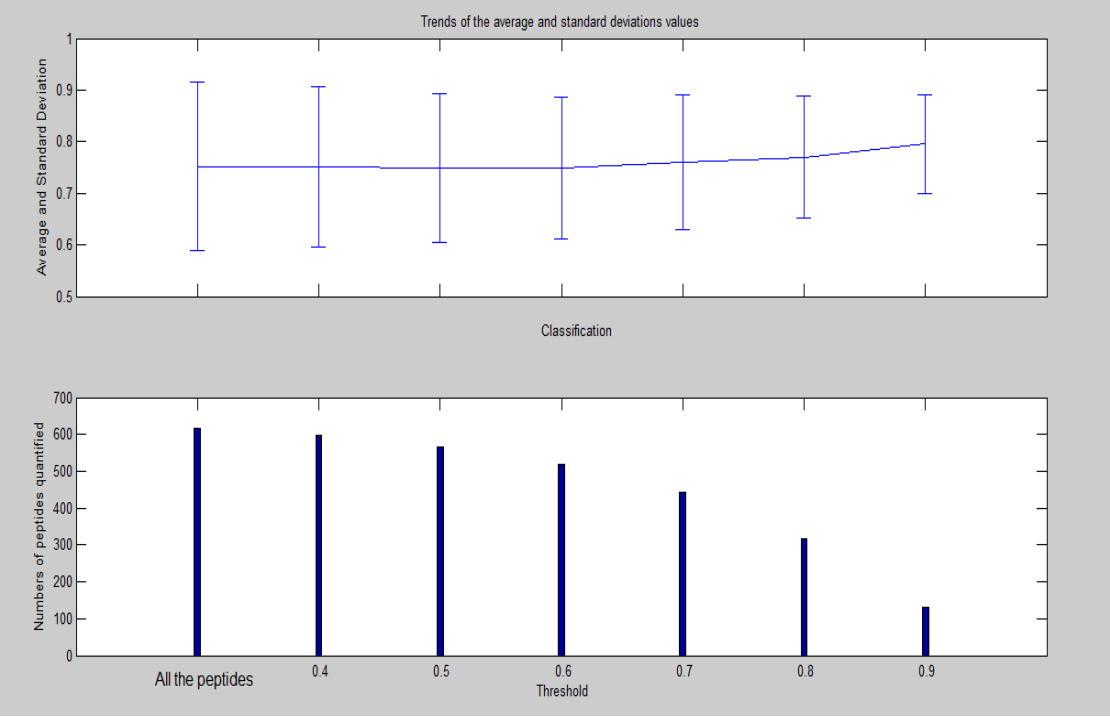


Fig.33 Classification and thresholding of the Medium/Light ratio in the Dataset SILAC02 1:1:1

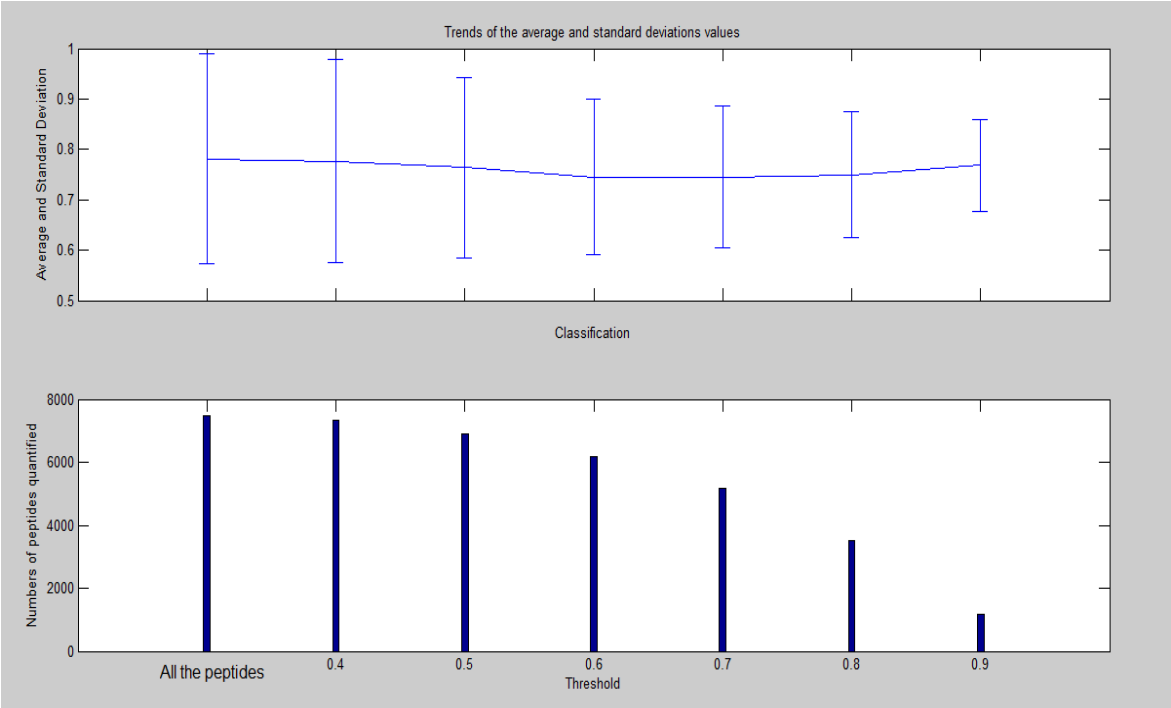


Fig.34 Classification and thresholding of the Medium/Light ratio in the Dataset SILAC03 1:1:1

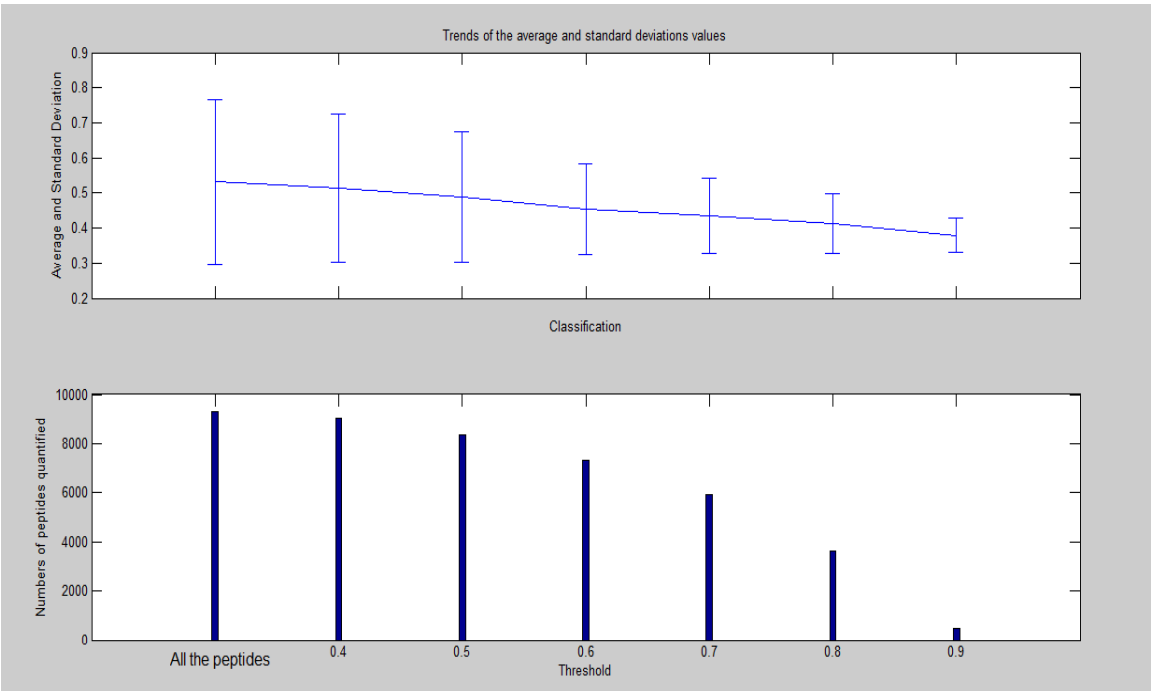


Fig.35 Classification and thresholding of the Medium/Light ratio in the Dataset SILAC02 10:5:1

Conclusions

We started this work with the aim of increasing the accuracy and the precision of the quantitation process of the state of art software (in particular the widely used MaxQuant). Specifically, we aim to solve the overlap problem, both in the co-eluting and in the overlapping case due to the kind of labeling (Light/Heavy overlap). As stated in the introduction: "*the aim of this work is the implementation of an algorithm able to improve the quantitation results obtained with the best state of the art algorithm, such as MaxQuant. In particular, we aim to improve the quantification of those peptides whose complexity, due to different kinds of overlap with other distributions, would lead to a significant poor quantification*."

The algorithm has been realized on the basis of two main ideas.

The selection of the scans from the elution area, using the theoretical distribution of the peptides to discard those scans whose shape is different or corrupted by noise (or overlap), is the first strategy used. As we have seen in the methods, such strategy allows us to use only that part of the signal which is not corrupted by noise. In this way, we have introduced a first filter in the information used to get the final ratio.

The second idea is to split the elution area of the peptide in several parts, both along the retention time axis and along the m/z axis, in order to compare different ratios and use only those which are supposed to be not overlapped with other peptides. Again, another important filter in order to choose only those parts of the area which are not corrupted.

Moreover, we have implemented an original idea based on the scoring of the quantification process at a peptide level. With this idea it has been possible to rank the peptides and to classify them on the basis of some important features that characterize the quality of the signal used for the quantitation. Such features, as seen, are the identification score, the number of sections used to get the final ratio, and the Pearson correlation between the shape of the peaks along the retention time axis. The final scoring of all the peptides of each dataset has been realized

using a single linear classifier, allowing a selection of peptides according to the quality of quantification. Such selection has the obvious drawback of a loss in terms of quantified proteins, which is the price to be paid to have an extremely accurate quantification.

In detail, compared with MaxQuant, the most used software in the proteomics field (due mainly to its free availability), our algorithm has obtained excellent results. For every dataset tested the precision of the quantification has been enhanced, and the number of the recovered proteins has been always comparable with that of MaxQuant.

Finally, the overlap issue about features of the same peptide (such as Light and Medium in the Dimethyl Dataset, the first one used and shown in the chapter about the results) has been successfully solved implementing the method published by Yoon et al [18]. Such method provides precise and accurate values. The results of the peptides quantified with such method show a significant enhancement in the precision, with smaller standard deviation of the results. Related with this overlap issue between features of the same peptide, we have also quantified the incidence of this problem in the whole proteome of the mouse.

In conclusion, at the end of this work, it is possible to say that our algorithm has successfully accomplished the task of improving the quantitation process of labeled peptides in a mass spectrometry experiment, both in the accuracy of the quantification results, and in the precision of the quantification, compared with the results obtained using MaxQuant; even the implementation of the solution proposed by Yoon for the overlap problem between elution area of different peptides, has given satisfying results, much better than those obtained with MaxQuant. The troublesome peptides badly quantified by MaxQuant, due to the overlap with some other elution areas, have been finally solved, as shown in the examples.

Finally, some future developments may include the possibility of using such algorithm even with other kind of labeling, trying to extend its usability with every kind of experimental protocol.

Bibliography

[1] *'Prepare for the deluge'* - Nature Biotechnology (2008) Volume: 26, Issue: 10, Pages: 1099

[2] http://www.geneontology.org/

[3] Ingvar Eidhammer, Kristian Flikka, Lennart Martens, Svein-Ole Mikalsen *'Computational Method for Mass Spectrometry'*; 2007 John Wiley & Sons, Ltd.

[4] *'Carl-Ove Andersson'*, Acta. Chem. Scand. **1958**, 12, 1353

[5] Makarov A. (2000). "*Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis".* Analytical Chemistry : AC **72** (6): 1156–62

[6] David K. Han, et al. *"Quantitative profiling of differentiation induced microsomal proteins isotope-coded affinity tags and mass spectronomy"* –Nat Biotechnol. 2001 October ; 19(10): 946–951.

[7] http://tools.proteomecenter.org/wiki/index.php?title=Software:XPRESS

[8] Jimmy K. Eng, Ashley L. McCormack, and John R. Yates, III "*An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database*". J Am Soc Mass Spectrom. 1994; 5 (11): 976–989.

[9] Brian D. Halligan, et al. *"ZoomQuant: An Application for the Quantitation of Stable Isotope Labeled Peptides"* J Am Soc Mass Spectrom. 2005 March ; 16(3): 302–306.

[10] Yergey J, Heller D, Hansen G, Cotter RJ, Fenselau C. *"Isotopic distributions in mass spectra of large molecules"*. Anal Chem 1983; 55:353–356.

[11] Yao X, Freas A, Ramirez J, Demirev PA, Fenselau C. *"Proteolytic 18O labeling for comparative proteomics: Model studies with two serotypes of adenovirus."* Anal Chem 2001; 73:2836–2842.

[12] Bellew et al.*"A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS"* Bioinformatics Vol. 22 no. 15 2006, pages 1902–1909

[13] Monroe, Tolic - *"VIPER an advanced software package to support high throughtput LC MS peptide identification"* Bioinformatics Vol. 23 no. 15 2007, pages 2021–2023

[14] Dasari et al. *"Quantification of Isotopically Overlapping Deamidated and ^{18}O Labeled Peptides Using Isotopic Envelope Mixture Modeling"* J Proteome Res. 2009 March 6; 8(3): 1263–1270

[15] Jurgen Cox and Matthias Mann *"MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification"* Nature Biotechnology 2008.

[16] http://maxquant.org/

[17] A.F. Maarten Altelaar, C.K.Frese et Al. *"Benchmarking stable isotope labeling based quantitative proteomics"* Nature Biotechnology 2012. /Ahead of print/

[18] Joo Young Yoon et al. *"Improved Quantitative Analysis of Mass Spectrometry using Quadratic Equations"* Journal of Proteome Research 2010, 9, 2775-2785

[19] Dirk Valkenborg, Ivy Jansen, and Tomasz Burzykowski *'A Model-Based Method for the Prediction of the Isotopic Distribution of Peptides'* J Am Soc Mass Spectrom. 2008, 19, 703–7127

[20] http://omics.pnl.gov/software/IPC.php

[21] Lukas Reiter, Oliver Rinner, Paola Picotti, Ruth Hüttenhain, Martin Beck, Mi-Youn Brusniak, Michael O Hengartner, Ruedi Aebersold: *'mProphet: automated data processing and statistical validation for large-scale SRM experiments'* Nature Methods, 2011 Vol. 8 n. 5

[22] http://omics.pnl.gov/software/ProteinDigestionSimulator.php

[23] S. Cappadona , J. Muñoz , Wim P.E. Spee, Teck Y. Low, S. Mohammed, Bas van Breukelen, Albert J.R. Heck *'Deconvolution of overlapping isotopic clusters improves quantification of stable isotope–labeled peptides'* Journal of Proteomics 74 (2011) 2204-2209

[24] M.J. MacCoss, C. Wu, H. Liu, R. Sadygov, J. R. Yates *'A Correlation Algorithm for the Automated Quantitative Analysis of Shotgun Proteomics Data'* Anal. Chem. 2003, 75, 6912 – 6921

[25] L.N. Mueller, Mi-Youn Brusniak, D.R.Mani, R.Aebersold *'An Assessment of Software Solutions for the Analysis of Mass Spectrometry Based Quantitative Proteomics Data'* Journal of proteme research 2008, 7, 51-61

[26] Shao Ong, L. Foster, Mattihas Mann, '*Mass spectrometric-based approaches in quantitative proteomics*' ScienceDirect, 2003, 124-130

[27] Pan S, Aebersold R. '*Quantitative proteomics by stable isotope labeling and mass spectrometry'*. Methods Mol Biol 2006;367:209–18.

[28] Tao WA, Aebersold R. *'Advances in quantitative proteomics via stable isotope tagging and mass spectrometry'*. Curr Opin Biotechnol 2003;14(1):110–8.

[29] Jacques Colinge, K.L. Bennet : '*Introduction to Computational Proteomics*' Plos Biology July 2007, Vol.3 Issue 7 e114

[30] R. Zhang, C. S. Sioma, S. Wang, and F. E. Regnier *'Fractionation of Isotopically Labeled Peptides in Quantitative Proteomics'* Anal. Chem. 2001, 73, 5142-5149

[31] Aebersold, R. & Mann, M. '*Mass spectrometry-based proteomics'*. Nature 422, 198–207 (2003).

Acknowledgements (ITA)

La parte dei ringraziamenti di solito è la parte più piacevole da scrivere, non solo perché non ci sono formule matematiche o astruse equazioni, ma perché è quella in cui ti volti indietro e guardi un po' fin dove sei arrivato, con quali compagni di viaggio hai condiviso le fatiche e le gioie del cammino.

Il primo sentito e doveroso ringraziamento va alla professoressa Linda Pattini, per avermi introdotto, grazie al Suo corso, nel meraviglioso mondo della Bioinformatica. La cosa buffa è che quasi per caso ho inserito il Suo corso nel piano di studi, al posto di un altro che proprio non mi interessava: allora non sapevo nemmeno bene cosa fosse questa "Bioinformatica". Inoltre, La ringrazio per la disponibilità dimostrata durante quest'anno nel seguirmi lungo questo percorso "proteomico", e per avermi infine dato la possibilità di trascorrere un paio di intense settimane all'università di Utrecht.

Il secondo grazie, o meglio "bedankt", lo rivolgo a Salvatore 'Salvo' Cappadona, che, lo posso assicurare, è stato un invidiabile punto di riferimento logistico, scientifico, tecnico e soprattutto umano, durante tutto il mio breve soggiorno a Utrecht. Con la sua solida competenza, ma anche con sorprendente ironia e disposizione, ha permesso alla tesi di compiere un enorme balzo di qualità. Un pensiero e un ringraziamento anche a sua moglie, Alessia, per la sua simpatia e la sua sorridente ospitalità, e alla loro bellissima bambina, Anna (ogni tanto mi sorprendo ancora, con un sorriso sulle labbra, a canticchiare tra me e me: "Hey, Jude"..!).

Restando in ambito Politecnico, ci tengo a ringraziare due persone, che probabilmente mai leggeranno queste pagine, ma che comunque sono state fondamentali per la mia formazione culturale in questi due anni di specialistica. Un sincero ringraziamento al Professor Sergio Rinaldi, docente del corso 'Dinamica dei Sistemi Complessi', e al Professor Pier Giorgio Righetti, docente del corso di 'Proteomica', che con il loro carisma e il loro modo stimolantissimo di fare lezione, sono state due preziose perle nel mio cammino.

Più di tutto, ciò che fa andare avanti nella vita sono gli affetti personali, le persone che si amano profondamente. Il primo pensiero va all'inesauribile supporto fornito dai miei genitori. So già che, pur non vedendo l'ora, adesso, di prendere il volo per una vita indipendente, tra qualche tempo non vedrò l'ora di tornare a casa a riabbracciarli. Un immenso grazie per il vostro amore e la vostra instancabile presenza.

Un'altra persona è stata, da ormai venticinque anni a questa parte, un importante punto di riferimento per la mia crescita. Sempre un passo avanti, la soluzione giusta ad ogni problema, il pensiero giusto per il momento giusto, è – ed è stato- un faro per la mia navigazione. Un grazie particolare a mio fratello, Marco.

Ed infine, il pensiero vola alla mia musa, Paola. Possa il nostro futuro essere felice e sereno, e restituirci tutti i momenti di gioia di cui il Poli ci ha temporaneamente privato.

Andrea