FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

Bioinformatics Master Thesis

Topics in Mass Sectrometry and Proteomics

Antonio Ortega Jiménez <ntoniohu@gmail.com>

Supervisors

Thomas Hamelryck <thamelry@gmail.com>
Mathias F. Gruber <mafg@novozymes.com>

Contents

Abbreviations

FT-ICR Fourier Transform Ion Cyclotron Resonance

HPLC High pressure liquid chromatography

IT Ion Trap

m/z mass charge ratio

MS Mass Spectrometry

MS/MS Tandem mass spectrometry

MS1 first tandem MS analyzer

MS2 second tandem MS analyzer

Q Quadrupole

TOF Time of Flight

Preface

Chapter 1

Introduction

1.1 Aminoacids and proteins

Proteins represent the last link in the central dogma of biology, where information encoded in DNA, is transcribed to RNA for posterior translation into proteins at the ribosome.

Proteins are made up of 20 basic units, called aminoacids. All aminoacids share a common chemical structure, where a carbon atom (C_α) is covalently bonded to a hydrogen atom, a carboxyl group, an amino group, and last but not least, a radical, also called side chain of the aminoacid (see figure **??**. The side chain differs between aminaocids and generates them from each other. A slight deviation from this pattern exists in proline, where the radical is bound to the nitrogen atom, making it an iminoacid. Even though the side chains are all different, they can be classified into four different groups: aliphatic, polar, positively charged and negatively charged (see figure 1.1).

Two aminoacids are joined together through the formation of a peptidic (covalent) bond between them. Such a linkage is formed by removal of the

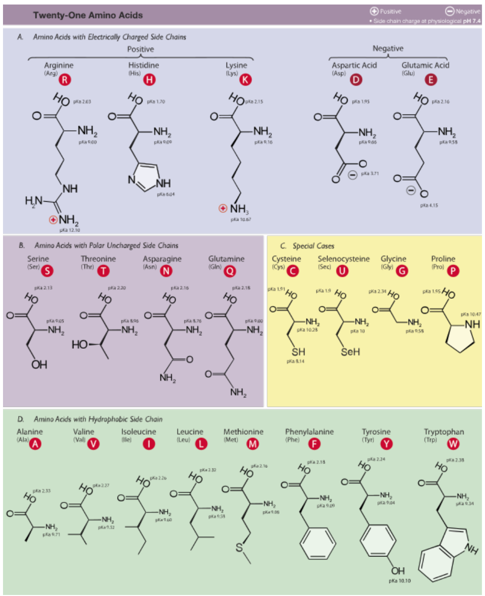


Figure 1.1

elements of water (dehydration) from the α-carboxyl group of one amino acid and the α-amino group of another. CITE LEHNINGER PAG 82. The remaining α-amino and α-carboxyl groups are available for linkage to other aminoacids, and in this way peptidic chains or peptides can be created.

While there are 20 basic units that constitute the majority of naturally observable proteins, their side chains can be modified both by physiological processes and by experimental procedures CITE. One frequent instance of such modifications is the oxidation of methionine

1.2 The protein-focused biotechnology industry

Proteins carry out most of the cell's molecular functions, they work as molecular agents that can perform an extremely wide range of tasks. The advent of biotechnology has sought to take advantage of this power, either by using proteins as present in natural conditions (wild type) or engineered by humans. This potential economic activity is carried out by several biotech companies, including Novozymes.

Novozymes A/S is a company whose line of business consists of the development of enzymatic products performing chemical transformations in different industrial processes. The application of these products, instead of conventional chemical-based solutions, has the advantage that they require less chemical substances, potentially simplifying industrial processes, reducing their costs and their environmental impact. Notorious examples of such applications include waste-water treatment, household care and the baking industry.

In order to place Novozymes ahead of its competitors, the company has utmost interest in developing protein research and several departments in the organization approach the study of proteins and their translation to the market through among other things the application of Mass Spectrometry (MS) and Proteomics.

1.3 Objectives of the Thesis

In line with the goal of making Novozymes more competitive, this project aimed at the following objectives:

1. Develop an open-source, Linux based and straightforward to deploy pipeline for the analysis of mass spectrometry data, starting at the output RAW files and ending in the interpretation of the results.

2. Evaluate this pipeline with a benchmark dataset to assess if the pipeline is able to reflect the biological phenomena captured in the data.

3. Implement a label-free quantification Bayesian-based method using Pymc3.

Chapter 2

Mass spectrometry and shotgun proteomics

Life systems consist of complex systems, meaning their behaviour cannot be easily explained by analyzing the individual elements alone. Moreover, they present multiple layers of complexity, given by the nature of the elements that make it up. The layer provided by proteins is one of them, and its study is called proteomics. It is a complex layer because thousands of different proteins can be present in a single cell at any time, and their exact composition and quantities constantly change, responding to the stimuli of the surrounding environment. The study of the protein-specific complexity is called proteomics. With proteomics, one endeavors to infer the protein composition of a sample, and eventually quantify its protein amounts.

It may be useful to divide the existing approaches into two types of paradigms: top-down and bottom-up. In the top-down paradigm, intact proteins are directly used for the analysis. In the bottom-up paradigm, the proteins are first cleaved into smaller parts, and these parts are then used for identification, characterization, and quantification. These smaller parts are called peptides. CITE 1.6.1 COMPUTATIONAL METHODS. Such peptides acquire physicochemical properties fitting the requirements of the downstream analytical methods, mainly mass spectrometry (MS), which performs the data acquisition. The bottom-up paradigm is most often used because peptides are much more suitable to analysis by mass spectrometry. The interested reader is directed to CITE 1.6.1 COMPUTATIONAL METHODS to learn why. The top-down paradigm will be ignored in the rest of the manuscript.

MS is performed by means of a mass spectrometer, an ensemble of pieces of equipment that can acquire mass measurements for plenty of sample components. A detailed explanation of the sample processing required prior to MS is given in section , while an overview on mass spectrometers is given in section . The result of the MS analysis is a dataset that, with adequate computational analysis tools, is enough to perform the inference steps required to gather knowledge about the original protein sample. These inference steps can be condensed to the peptide and protein inference problems, explained in section . A third computational problem needs to be solved if quantitative, and not just qualitative information, is to be gained from the experiment. This is the quantification problem, explained in section CITE.

A summary of the bottom-up approach MS analytical pipeline follows.

2.1 Sample processing

An MS experiment starts with the generation of protein mixture samples. They are first separated in order to sort the proteins via physicochemical criteria. This is most frequently carried out via SDS-PAGE based on mass or isoelectric point. Once the electrophoresis is completed, protein bands can be excised from the gel. Each band will contain a subset, or even only one of the proteins originally available, thus making the downstream analysis simpler REFERENCE COMPUTATIONAL METHODS CHAPTER 2

The reduced complexity protein mix is extracted in a denaturalized state and subjected to enzymatic digestion with specific enzymes, that can cut the chain of aminoacids following a predictable pattern. The enzyme most frequently used for this is Trypsin, which cuts peptidic bonds whenever a positively charged residue, either Lysine (K) or Arginine (R), lies on the carboxyl side of the peptidic bond. Even though enzymes are very specific, the cleaving process is far from perfect, as there could be: COMPUTATIONAL METHODS 3.2

1. Missed cleavages
2. Unsuspected cleavages during the maturation/life cycle of the protein.
3. Unexpected cleavages occurring either in the wet-lab procedure of the proteolytic treatment.
4. Naturally occurring, intentionally or unintentionally induced chemical modifications.

Item 1 can happen due to steric inaccesability or the presence of specific aminoacids that can weaken the enzyme's function. This is the case of Trypsin whenever the residue on the other side of the peptidic bond is Proline. This variability, though limited, needs to be taken care of in downstream analysis, as it could introduce biases in peptide observability. The other

The result of this process is a mix of peptides following a length distribution given by the cleavage sites frequency and each protein's aminoacidic composition. For Trypsin, the average peptide length is 10 residues, as roughly 1/10 residues are either R or K. As explained in 2.2.3, this length distribution is fitted to the resolution of the MS analyzer, thus optimizing the throughput of the method. An overview over mass spectrometers follows.

2.2 The mass spectrometer

The mass spectrometer consists of three main parts: an ion source, a mass analyzer, and a detector SEE FIGURE.

2.2.1 The ion source

All mass spectrometers exploit the physical properties of mass and electric charge exhibited by the analyzed components. Ionization of the analytes is absolutely essential prior to any measurement, as analytes left uncharged will be unobservable to the equipment. This step is performed in the ion source CITE 5.1 COMPUTATIONAL METHODS. The most frequent ionization methods in proteomics are Matrix-Assisted Laser Desorption-Ionization (MALDI) and Electro Spray Ionization (ESI) CITE. Most peptides ionized by MALDI will acquire a single charge, whereas ESI can provide multiple charges (+2, +3, etc) too. Thus, the charge exhibited by an ion is not obvious when produced via ESI. Moreover, ESI can be run online with the right robotic equipment, while MALDI demands waiting time for vacuum generation. Finally, due to the chemical nature of the matrix components, MALDI ionizes more easily peptides containing aminoacids featuring aromatic rings (PYW), thus introducing a bias. Bias in ESI is less predictable. This is known as the competitive ionization problem. REF ALL THIS

The acquired charge yields a mass/charge (m/z) ratio, a property that can be applied in the downstream component separation and measurement steps.

2.2.2 The mass analyzer

The plethora of ion separation methods is reflected upon the range of different analyzers available, mainly time of flight (TOF), Ion trap (IT) and quadrupole (Q). These apply different principles to perform the same task: separation (analysis) of the ion mix by the m/z ratio.

Moreover, two other analyzers exist which combine mass analysis with intensity measurement. These are Fourier Transform Ion Cyclotron Resonance (FT-ICR) and Orbitrap. They both register cylotron resonance frequencies that are Fourier transformed into the spectrum space. Remarkably, FT-ICR exhibits great resolving power, at the cost of high maintenance costs and difficult operability.

2.2.3 The detector

Detectors measure the intensity of an incoming ion signal. The ion's m/z ratio is known thanks to the previous mass analysis step. Performed for enough m/z ratios, the detector can produce a MS spectrum, which shows the intensity of an ion signal over an m/z range. Some topics in signal detection in MS need to be discussed.

On the one hand, the precision of the signal measurement is given by its mass resolution It is conventionally defined as the closest distinguishable separation between two peaks of equal height and width [**?**]. The resolution decreases as the m/z ratio increases because small increments in the m/z ratio become negligible at high m/z ratios. This is one of the reasons why proteins are better fit for analysis when digested into peptides, as m/z are reduced, thus increasing the mass resolution.

On the other hand, due to the natural occurrence of isotopes, particularly ^{13}C, the same peptide will induce the measurement of several signals with very close m/z values. They constitute the isotopic envelope of the ion SEE FIGURE, and represent the signal created by peptides containing an increasing number of ^{13}C atoms. Every time a ^{12}C is replaced by ^{13}C, the mass increases by 1 Da. Even though the natural abundance is 1.1 %, the sheer number of carbon atoms in a peptide makes it likely that at least one or even more carbon atoms are ^{13}C, eventually driving the pure ^{12}C signal to comparatively small intensity values, and down to intensities below the background noise. Such event can pe problematic if it entails that the 1 ^{13}C peak is confused for the completely ^{12}C peak.

The resolution achieved by modern equipment allows for the distinction of each individual signal in most isotopic envelopes. Remarkably, the separation across peaks in the envelope can be used to infer the charge of the peptide, as increases of 1 Da at charge 1 will induce a separation of 1 m/z, while at charge 2 it will be $1/2 = 0.5$ m/z, at 3 $1/3 = 0.33$ m/z, and so on.

It is up to the MS technician to decide on the best pieces of equipment according to the particularities of the dataset.

2.3 Mass spectrometry workflows

The MS workflow now diverges based on the simplicity of the original protein sample. When it consisted of a single protein, Peptide Mass Fingerprinting (PMF) is used, otherwise tandem MS (MS/MS) shall be performed.

2.3.1 Peptide Mass Fingerprint (PMF)

If the original sample was known to contain a single protein, PMF, or *protein-centric* proteomics, is conducted. In PMF, the mixture of peptides can be already transferred to the spectrometer, where a a spectrum containing a peak for every m/z ratio present in the ionized peptide mix will be recorded. Thus, spectra generated this way can be considered a pattern, or fingerprint, of the peptides making up the original protein. Therefore, the spectra are yield information that can be used to identify the original protein.

2.3.2 High pressure liquid chromatography-Tandem MS

HPLC

If presented with the problem of analyzing a mixture of proteins, the capacities of mass spectrometers are easily overwhelmed by a too complex mixture, resulting in the analysis of only a minor part of the total protein of the sample. This can be surmounted by splitting the initial sample into fractions, and using a series, or tandem, of spectrometers in the analysis. The spectrometers are used to analyze each obtained fraction separately, using different schemes. Fractionation is usually achieved by different methods of separation CITE 1.7, most commonly via High pressure liquid chromatography (HPLC) methods.

HPLC methods work by loading the peptide mix in a column containing a stationary and a solid phase. These phases create an environment where peptides interact differently based on their physico-chemical properties, set by the nature of the phases. The output of the column, called elute, will consist of subsets or fractions of peptides leaving the column at different retention times. Therefore, the input to the machine will consist of a simplified mix of peptides at a time. Notably, the same peptide could elute in several contiguous fractions.

The two most common fractionation methods in proteomics are reverse phase chromatography (separating on hydrophobicity) and strong cation exchange chromatography (separating on isoelectric point) CITE 4.2 computational methods.

The tandem MS MS/MS) analysis starts when the peptide mix accesses the analyzer. It is ionized in the ion source and enters the first mass spectrometer. While different ways of handling the peptides are available, we will focus on the product ion scan method. In this protocol, the first analyzer is used to select ionized peptides within a narrow m/z window.

Fragmentation

Peptides passing this first scan are then subjected to fragmentation, most often via (I) collision-induced (CID) or (II) electron-induced (EID) dissociation. This way, peptides are further processed into smaller fragments resulting from the breakage of either the peptidic bond, the C_α-CO or the C_α-NH bonds

In CID, peptides enter a collision cell containing an inert gas. Given enough kinetic energy, hits of ionized peptides and the gas will trigger the fragmentation of the peptide into smaller units. PAG 123 COMPUTATIONAL METHODS. The most frequently occurring fragment types are the b and y ions. pag 134 computational methods. The produced fragmetns then enter the second analyzer, where a m/z spectrum of the fragments is recorded. Thus, unlike in PMF, where the spectrum recorded reflects the m/z ratios acquired by the protein peptides cleaved by the enzyme, tandem MS spectra on product ion scan mode record the m/z ratio of the fragments produced by an ionized peptide with a given m/z ratio. The m/z ratio of this precursor ion is changed during the run, thus, multiple spectra are obtained where PMF would create only one.

2.4 Spectra processing: search engines

A search engine can be used to map the recorded pattern of m/z ratios to a protein entry in a database. This task is performed by *in silico* cleavage of each sequence entry in the database based on the specefic cleavage pattern of the enzyme used, coupled with the simulation of the expected spectrum based on the expected peptides.

Given the stochastic nature of the cleavage and spectra recording process, the resulting spectra exhibit variability manifested in missing peaks or spurious ones. Furthermore, the peptide to spectrum matching (PSM) process against a sufficiently big database can, at random, return wrong matches. This translates to the obtention of multiple matches , of which one, if any, will be correct. Therefore, the lists of matches need to be somehow ranked. The issue is addressed by search engines through the usage of scoring systems that measure the goodness of the match. Assuming the correct protein is present in the database, a good system should give the protein the best score. This way, proteins can be identified.

Two steps in protein identification can be distinguished:

1. Peptide inference: infer the peptides present in the sample.
2. Protein inferece: based on the infered peptides, infer what proteins generated them.

Both are taken care of by the search engine. Multiple search engines exist that implement different matching and scoring algorithms.

2.4.1 X!Tandem

X!Tandem was one of the first open source search engines in mass spectrometry. It produces a score based on the dot product between the theoretical and the experimental tandem mass spectra CITE Overview of Tandem Mass Spectrometry (MS/MS) Database Search Algorithms. The scores assigned to wrong matches are assumed to follow a hypergeometric distribution, allowing the program to extrapolate $E(s)$ (expected number of wrong matches at a given score) for any score.

2.5 Results validation

The scoring system implemented by the search engines can be used to filter and validate the results. A basic common filter is the false discovery rate (FDR), usually set to 1%, indicating that 1 out of a hundred filtered matches are expected to be false positives. The most commonly used method to compute the FDR is the target-decoy search. The search engine tries to match the same spectra against a decoy database, usually generated by reversing the sequences present in the original database (target). All matches to the decoy will be regarded as wrong. Thanks to the fact that the basic properties of the decoy (size, composition, etc) are identical to the target, whenever a mistake is made, it is as likely to happen in both databases CITE COMPOMICS TUTORIAL 1.5, thus the number of matches in the decoy provides an accurate estimate of the number of random matches, or false positives, against the target database (n_{fp}). Together with the number of PSMs passing a threshold score ($n_{tp} + n_{tp}$), the FDR can be computed using the formula below.

$$FDR = \frac{n_{fp}}{n_{tp} + n_{tp}}$$

The minimal FDR at which a given PSM is considered a positive match constitutes the PSM's q-value

$$q(PSM_i) = \min FDR \forall PSM_i \in Positives$$

Finally, a posterior error probability (PEP) can be defined as the probability of the match being random ($P(s_i|T_i = 0)$). This can only be done if a statistical model describing the distribution of scores for correct and wrong matches is fit to the dataset.

2.6 Protein quantification

CONTINUE maybe add reason for the project.

Chapter 3

Materials and Methods

Chapter 4

Results and Discussion

Chapter 5

Conclusion

5.1 Overview of proteomics analysis software

5.1.1 MaxQuant

One of the most used program for the quantification of label free proteins is the MaxQuant suite [**?**]. It can be downloaded for free and is developed by the Max Planck institute in Germany. It consists of a user friendly GUI that provides the most needed steps in a proteomics pipeline.

MaxQuant has been successfully adopted by the scientific community due to its ease of use and a comprehensive pipeline. A Google group and a tutorial are available [1] [2].

The main processing steps incorporated in MaxQuant consist of:

- Read MS spectra files in the .RAW format, the closed format produced by ThermoScientific MS analyzers.

[1] http://www.coxdocs.org/doku.php?id=maxquant:viewer:tutorial

[2] https://www.youtube.com/watch?v=_AJHFHi5CxM

- Contains the Andromeda search engine [?] for matching of MS2 spectra to a proteome database.
- Andromeda supports configuring aminoacid modifications and decoy search for FDR (false discovery rate) estimation.
- Increase peptide identifications by performing match between runs (MBR), which aims at transferring peptide to spectrum matches (PSMs) across replicate runs, based on precursor mass and retention times.
- Filters Andromeda results to provide a list of inferred peptides and proteins with a cutoff FDR value.
- Supports label-based and label-free quantification.
- Its output can be passed to the Perseus software for data visualization.

However, MaxQuant also suffers from some problems:

1. MaxQuant only runs on Microsoft Windows, impeding its integration in automated pipelines on cluster environments [?]. The processing steps cannot easily be fine tuned or exchanged with other pipelines. For example, only the Andromeda search engine is supported. This is a serious drawback, as the integration of results from several search engines, like MSGF+ [?] or MS-Amanda [?] has been shown to further improve results [?].
2. It does not provide a command line interface (CLI), thus all analyses must be performed through the GUI. This hinders reproducibility and scalability.
3. While it is supported in Linux through Mono, the user experience is best in Windows. As Linux is by far the most used OS in Bioinformatics, this implies that an additional OS is required to get the best

experience with MaxQuant.

4. Lacks a well written official documentation, as most information is made available only through talks published on Youtube or third party tutorials.

5.1.2 Compomics group

The Compomics group at Ghent University [3] provides high quality analysis software for the identification and integration of results through the SearchGUI, PeptideShaker and moFF (modest Feature Finder) programs [?] [?] [?].

- SearchGUI provides a common interface to a range of search engines so that multiple engines can be used in a straightforward manner.
- PeptideShaker reads the SearchGUI output and performs quality control, gene set enrichment analysis (GSEA), and implements multiple validation filters to provide the best results.
- moFF reads from PeptideShaker output and implements a match between runs (MBR) model analogous to that in MaxQuant, to increase the number of spectra that are matched to a peptide. Furthermore, it summarises the MS1 peaks into refined features, making downstream analyses more sensitive.

SearchGUI and PeptideShaker are very well integrated and documented [4]. Both provide not just beautiful GUIs, but also a comprehensive and extremely well documented CLI [5] [6]. They easily enable the development of analytical pipelines finely fitted to each use case.

[3] https://compomics.com/

[4] https://compomics.com/bioinformatics-for-proteomics/

[5] https://github.com/compomics/searchgui/wiki/SearchCLI

[6] https://github.com/compomics/peptide-shaker/wiki/PeptideShakerCLI

moFF is less documented and integrated in the workflow. Furthermore, it does not offer a GUI yet. The group is currently working to better integrate moFF and PeptideShaker [7].

The main issue in the Compomics suite of tools is the lack of a robust label-free quantification step, as they currently recommend the export of results to a third party tool. Only the spectral count based methods emPAi (Exponentially Modified Protein Abundance Index) and NSAF (Normalized Spectral Abundance Factor), which are not robust [**?**], are supported.

5.1.3 Other suites

Other software suites like ProteomeDiscoverer, Progenesis QI, GeneData Expressionist, etc, are available. They are not open, requiring a license to function. Furthermore, they are GUI oriented and make use of different data formats thus, making exchange across pipelines extremely difficult. The OpenMS [**?**] and Trans Proteomic Pipeline (TPP) [**?**] suites are open, but they also suffer from data exchangeability and lack comprehensive documentations.

5.2 Label-free quantification

5.2.1 MaxLFQ

The label-free quantification engine implemented in MaxQuant is MaxLFQ [**?**]. It performs 2 minimisation steps to infer protein quantities from extracted ion currents (XIC), as stored in the spectra files.

[7] https://groups.google.com/forum/#!topic/peptide-shaker/Lqe7lYKLcHI

Fractionated XIC aggreagation

First, due to the prefrationation of samples during the upstream mass spectrometry (MS) analysis, XIC signals for each peptide are distributed along multiple fractions. MaxLFQ summarises the XIC from several fractions for each peptide and sample into a single peptide intensity by minimising the sum of the square of the logarithm of the intensity between samples via Levenberg–Marquardt optimization (see figure 5.1 for a graphical explanation).

The intensity of peptide P in sample A is defined as a weighted average of P 's XIC signals across fractions $\{1..k\}$, where the weights are the normalization factors that the minimisation algorithm seeks to find.

$$I_{P,A}(\mathrm{N}) = \sum_{j=1}^{k} \mathrm{N}_{P,A,j} \times \mathrm{XIC}_{P,A,j} \tag{5.1}$$

The minimisation assumes that the best normalization factors minimize the sum of squared logarithm pairwise-ratios H.

$$H_P(N) = \sum_{a,b} \left| \frac{I_{P,a}(N)}{I_{P,b}(N)} \right|^2 \tag{5.2}$$

$$H(N) = \sum_{p} H_p(N) \tag{5.3}$$

where a, b iterates over all possible pair-wise combinations of samples where P was detected.

The peptide intensities can be computed from the estimated normalization factors and the observed XIC signals.

Protein intensity inference

Second, once the peptide intensities have been summarised, the intensity of the proteins that the peptides originated from are inferred. This is achieved

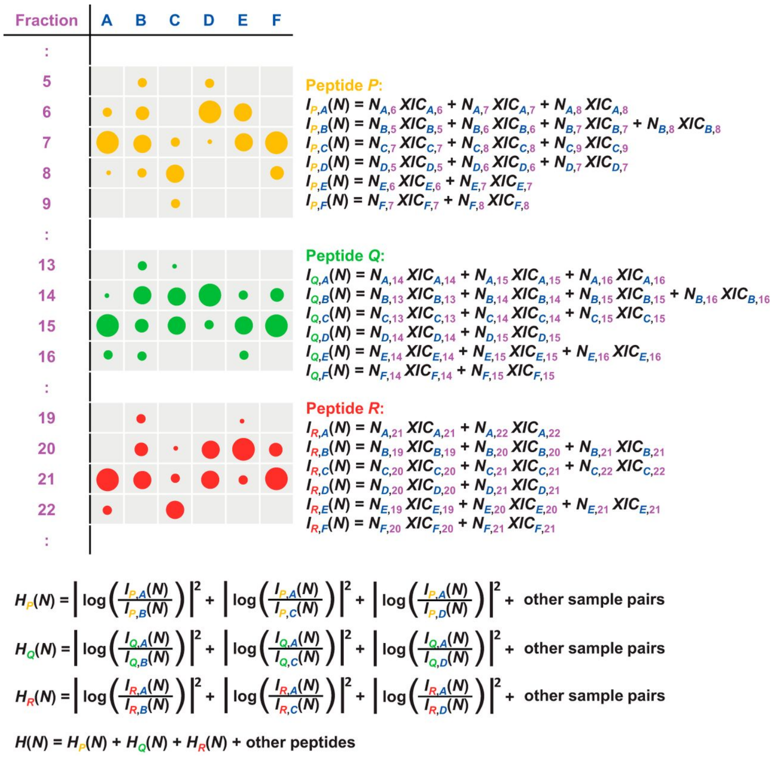

Fraction
A
B
C
D
E
F
5
6
7
8
9
13
14
15
16
19
20
21
22
Peptide P:
$I_{P,A}(N) = N_{A,6}\,XIC_{A,6} + N_{A,7}\,XIC_{A,7} + N_{A,8}\,XIC_{A,8}$
$I_{P,B}(N) = N_{B,5}\,XIC_{B,5} + N_{B,6}\,XIC_{B,6} + N_{B,7}\,XIC_{B,7} + N_{B,8}\,XIC_{B,8}$
$I_{P,C}(N) = N_{C,7}\,XIC_{C,7} + N_{C,8}\,XIC_{C,8} + N_{C,9}\,XIC_{C,9}$
$I_{P,D}(N) = N_{D,5}\,XIC_{D,5} + N_{D,6}\,XIC_{D,6} + N_{D,7}\,XIC_{D,7}$
$I_{P,E}(N) = N_{E,6}\,XIC_{E,6} + N_{E,7}\,XIC_{E,7}$
$I_{P,F}(N) = N_{F,7}\,XIC_{F,7} + N_{F,8}\,XIC_{F,8}$
Peptide Q:
$I_{Q,A}(N) = N_{A,14}\,XIC_{A,14} + N_{A,15}\,XIC_{A,15} + N_{A,16}\,XIC_{A,16}$
$I_{Q,B}(N) = N_{B,13}\,XIC_{B,13} + N_{B,14}\,XIC_{B,14} + N_{B,15}\,XIC_{B,15} + N_{B,16}\,XIC_{B,16}$
$I_{Q,C}(N) = N_{C,13}\,XIC_{C,13} + N_{C,14}\,XIC_{C,14} + N_{C,15}\,XIC_{C,15}$
$I_{Q,D}(N) = N_{D,14}\,XIC_{D,14} + N_{D,15}\,XIC_{D,15}$
$I_{Q,E}(N) = N_{E,14}\,XIC_{E,14} + N_{E,15}\,XIC_{E,15} + N_{E,16}\,XIC_{E,16}$
$I_{Q,F}(N) = N_{F,14}\,XIC_{F,14} + N_{F,15}\,XIC_{F,15}$
Peptide R:
$I_{R,A}(N) = N_{A,21}\,XIC_{A,21} + N_{A,22}\,XIC_{A,22}$
$I_{R,B}(N) = N_{B,19}\,XIC_{B,19} + N_{B,20}\,XIC_{B,20} + N_{B,21}\,XIC_{B,21}$
$I_{R,C}(N) = N_{C,20}\,XIC_{C,20} + N_{C,21}\,XIC_{C,21} + N_{C,22}\,XIC_{C,22}$
$I_{R,D}(N) = N_{D,20}\,XIC_{D,20} + N_{D,21}\,XIC_{D,21}$
$I_{R,E}(N) = N_{E,19}\,XIC_{E,19} + N_{E,20}\,XIC_{E,20} + N_{E,21}\,XIC_{E,21}$
$I_{R,F}(N) = N_{F,20}\,XIC_{F,20} + N_{F,21}\,XIC_{F,21}$
$H_P(N) = \left|\log\left(\frac{I_{P,A}(N)}{I_{P,B}(N)}\right)\right|^2 + \left|\log\left(\frac{I_{P,A}(N)}{I_{P,C}(N)}\right)\right|^2 + \left|\log\left(\frac{I_{P,A}(N)}{I_{P,D}(N)}\right)\right|^2 +$ other sample pairs
$H_Q(N) = \left|\log\left(\frac{I_{Q,A}(N)}{I_{Q,B}(N)}\right)\right|^2 + \left|\log\left(\frac{I_{Q,A}(N)}{I_{Q,C}(N)}\right)\right|^2 + \left|\log\left(\frac{I_{Q,A}(N)}{I_{Q,D}(N)}\right)\right|^2 +$ other sample pairs
$H_R(N) = \left|\log\left(\frac{I_{R,A}(N)}{I_{R,B}(N)}\right)\right|^2 + \left|\log\left(\frac{I_{R,A}(N)}{I_{R,C}(N)}\right)\right|^2 + \left|\log\left(\frac{I_{R,A}(N)}{I_{R,D}(N)}\right)\right|^2 +$ other sample pairs
$H(N) = H_P(N) + H_Q(N) + H_R(N) +$ other peptides

Figure 5.1 Taken from [**?**]

in three steps (see figure 5.2 for a graphical explanation).

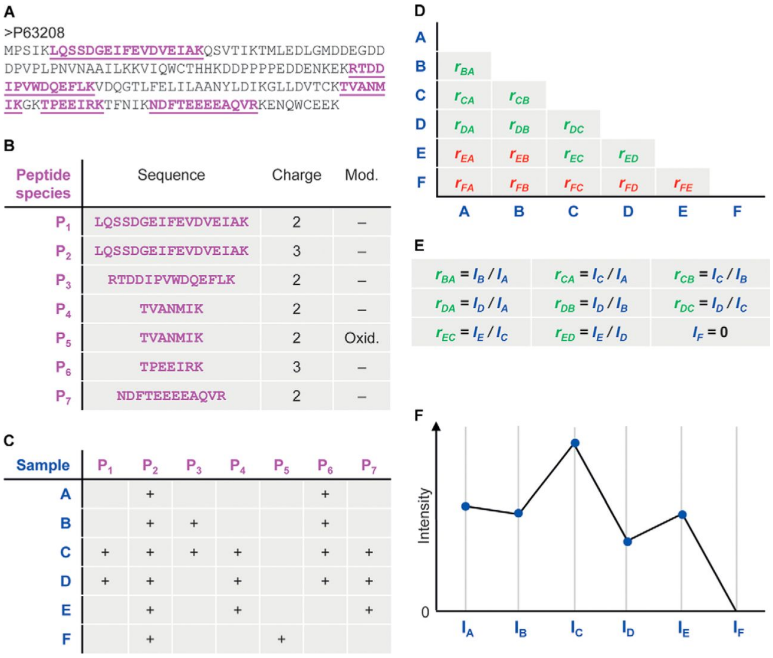
1. Peptide intensity ratios are computed for all possible pairwise combinations by dividing the normalized intensities obtained in the previous step.

2. For every inferred protein A, its protein ratio $r_{A,a,b}$ for the pair of samples a,b is set to the median intensity of its children peptide ratio in samples a,b. The median is selected as a summarising statistic to protect from outliers. A minimal number of non-zero intensity peptides are required for the median to be valid, usually 2. Otherwise, the protein ratio for the pair of samples is set to 0.

3. The correct protein intensities are assumed to minimise the sum of the squared difference of the logarithm of the protein ratios and log-

arithm of the protein intensities I.

$$\sum_{a,b} \left(\log\left(r_{A,a,b}\right) - \log\left(I_{A,a}\right) + \log\left(I_{A,b}\right)\right)^2 \tag{5.4}$$



A

>P63208
MPSIKLQSSDGEIFEVDVEIAKQSVTIKTMLEDLGMDDEGDD
DPVPLPNVNAAILKKVIQWCTHHKDDPPPPEDDENKEKRTDD
IPVWDQEFLKVDQGTLFELILAANYLDIKGLLDVTCKTVANM
IKGKTPEEIRKTFNIKNDFTEEEEAQVRKENQWCEEK

B

Peptide species	Sequence	Charge	Mod.
P_1	LQSSDGEIFEVDVEIAK	2	–
P_2	LQSSDGEIFEVDVEIAK	3	–
P_3	RTDDIPVWDQEFLK	2	–
P_4	TVANMIK	2	–
P_5	TVANMIK	2	Oxid.
P_6	TPEEIRK	3	–
P_7	NDFTEEEEAQVR	2	–

C

Sample	P_1	P_2	P_3	P_4	P_5	P_6	P_7
A		+				+	
B		+	+			+	
C	+	+	+	+		+	+
D	+	+		+		+	+
E		+		+			+
F		+			+		

D

A						
B	r_{BA}					
C	r_{CA}	r_{CB}				
D	r_{DA}	r_{DB}	r_{DC}			
E	r_{EA}	r_{EB}	r_{EC}	r_{ED}		
F	r_{FA}	r_{FB}	r_{FC}	r_{FD}	r_{FE}	
	A	B	C	D	E	F

E

$r_{BA} = I_B / I_A$	$r_{CA} = I_C / I_A$	$r_{CB} = I_C / I_B$
$r_{DA} = I_D / I_A$	$r_{DB} = I_D / I_B$	$r_{DC} = I_D / I_C$
$r_{EC} = I_E / I_C$	$r_{ED} = I_E / I_D$	$I_F = 0$

Figure 5.2 Taken from [?]

For the three most intense tryptic peptides, the signal per mole of protein was shown to be constant within a coefficient of variation of $\pm 10\%$ [?]. Thus, protein intensities are linearly proportional to protein quantities and protein quantity ratios can be computed from protein intensity ratios. These values are reported in the MaxQuant output in the `proteinGroups.txt` file under columns named *LFQ Intensity*.

Drawbacks of the model

The LFQ model provides a powerful method to infer protein quantities from a label-free experiment. However, it makes use of point estimates in different steps of the process, namely (1) during peptide intensity aggregation, where the Levenberg-Marquardt minimum is selected (2) during protein ratio estimation, where the peptide median is selected and (3) during protein intensity estimation, where the least squares minimum is selected. These point estimation bottlenecks, specially (2), discard increasing amounts of data that could otherwise be used for more accurate results. Moreover, not just quantities, but also the uncertainty behind them, could also be provided. The only way of MaxQuant to provide uncertainty measurements is the number of peptides supporting the quantification measurement.

5.2.2 Bayesian model

Bayesian statistics provides probability measurements for observed data assuming underlying mathematical models. A model for protein quantification based on the MS1 intensities or XICs could be developed to assess both quantities and uncertainties behind the estimated quantities.

While MS1 intensity measurements are dependent upon the underlying protein quantities, other factors influence the final measurement, thus adding noise and distorting the results. Besides the actual quantity, two more factors can be distinguished:

- **Sequence derived factors**: some peptides are easier to cleave for the cleaving enzyme (Trypsin, etc) than others. Moreover, the sequence of the peptide could influence the final measurement detected in the analyzer, by having different elution dynamics in the column or the analyzer.
- **Random noise**: the stochastic processes intrinsic to MS1 measurements could also have an impact.

The model would take as input:

1. **Features extracted from the precursor sequence**, including the surrounding aminoacids in the original protein. They can be used to model the sequenced derived factors.
2. **Observed MS1 precursor intensities/XICs**.

and return as output a protein quantification value for each protein, together with a measurement of the uncertainty behind it.

5.2.3 Data

The MaxLFQ paper dataset is available [8]. The authors submitted the RAW files (>50 GB) and the search files containing the final results. They can be used to benchmark any new quantification tool or method, for example, it has been used to benchmark third party quantification tools like StPeter from the TPP [**?**].

The dataset contains the results of the quantification of proteins in 2 different proteomes (*E. coli* and *Homo sapiens*) from 6 samples organized in 2 conditions with 3 replicates each. In the first condition (H), both proteomes were mixed in a 1:1 ratio, while in the second condition (L), the *E. coli* proteome had 3 times more contribution to the mix (3:1). Thus, the fold change of protein quantities across conditions should be 1 for human proteins and 3 for bacterial proteins.

A list of the protein groups identified that could be mapped unambiguously to each species was made available [9]. Thus, a dataset of thousands of proteins of known ratios between conditions, producing tens of thousands of peptides with known sequence, MS1 intensity/XIC and parent

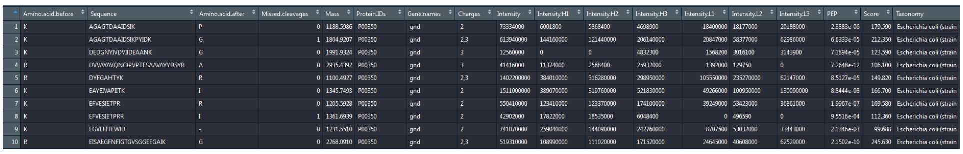
[8] https://www.ebi.ac.uk/pride/archive/projects/PXD000279

[9] http://www.mcponline.org/content/13/9/2513/suppl/DC1

protein can be built from the publication, and used to train new quantification methods (figure 5.3).

	Taxonomy	Peptides	Proteins
1	Escherichia coli (strain K12)	14483	1556
2	Homo sapiens	32647	3444
	Total	47130	5000

Table 5.1 Summary table of the dataset



	Amino.acid.before	Sequence	Amino.acid.after	Missed.cleavages	Mass	Protein.IDs	Gene.names	Charges	Intensity	Intensity.H1	Intensity.H2	Intensity.H3	Intensity.L1	Intensity.L2	Intensity.L3	PEP	Score	Taxonomy
1	K	AGAGTDAAIDSIK	P	0	1188.5986	P00350	gnd	2	73334000	6001800	5868400	4698900	18400000	18177000	20188000	2.3883e-06	179.590	Escherichia coli (strain
2	K	AGAGTDAAIDSIKPYIDK	G	1	1804.9207	P00350	gnd	2,3	613940000	144160000	121440000	206140000	20847000	58377000	62986000	6.6333e-05	212.350	Escherichia coli (strain
3	K	DEDGNYIVDVIIDEAANK	G	0	1991.9324	P00350	gnd	3	12560000	0	0	4832300	1568200	3016100	3143900	7.1894e-05	123.590	Escherichia coli (strain
4	R	DVVAYAVQNGIPVPTFSAAVAYYDSYR	A	0	2935.4392	P00350	gnd	3	41416000	11374000	2588400	25932000	1392000	129750	0	7.2648e-12	106.100	Escherichia coli (strain
5	R	DYFGAHTYK	R	0	1100.4927	P00350	gnd	2,3	1402200000	384010000	316280000	298950000	105550000	235270000	62147000	8.5127e-05	149.820	Escherichia coli (strain
6	K	EAYEIVAPIITK	I	0	1345.7493	P00350	gnd	2	1511000000	389070000	319760000	521830000	49266000	100950000	130090000	8.8444e-08	166.700	Escherichia coli (strain
7	K	EFVESIETPR	R	0	1205.5928	P00350	gnd	2	550410000	123410000	123370000	174100000	39249000	53423000	36861000	1.9967e-07	169.580	Escherichia coli (strain
8	K	EFVESIETPRR	I	1	1361.6939	P00350	gnd	2	42902000	17822000	18535000	6048400	0	496590	0	9.5516e-04	112.360	Escherichia coli (strain
9	K	EGVFHTEWID	-	0	1231.5510	P00350	gnd	2	741070000	259040000	144090000	242760000	8707500	53032000	33443000	2.1346e-03	99.688	Escherichia coli (strain
10	R	EISAEGFNFIGTGVSGGEEGAIK	G	0	2268.0910	P00350	gnd	2,3	519310000	108990000	111020000	171520000	24645000	40608000	62529000	2.1502e-10	245.630	Escherichia coli (strain

Figure 5.3 Extract of the available dataset. Total of 47130 entries, one per peptide, for a global of 5000 proteins.

This compiled dataset can be downloaded here https://mega.nz/#!AgcTgJYa!w9DoAKYRc6u-SaRy_UIMz3aileUHXoaWgrxf-UycqiQ

5.3 Extension of pipelines

The Compomics suite of programmes provides the best documented and accesible set of tools for proteomics analysis, but still lacks proper quantification tools. The implementation of a MS1 intensity-based downstream quantification tool compatible with the output of the Compomics tools would yield a free Linux compatible, CLI supported, complete proteomics pipeline featuring a robust quantification method.