Traitement du signal et images LC/MS pour la recherche de biomarqueurs

Sébastien Li-Thiao-Té

THÈSE DE DOCTORAT
DE L'ÉCOLE NORMALE SUPÉRIEURE DE CACHAN

Présentée par

Monsieur Li-Thiao-Té Sébastien

pour obtenir le grade de

DOCTEUR DE L'ÉCOLE NORMALE SUPÉRIEURE DE CACHAN

Domaine :
MATHÉMATIQUES APPLIQUÉES

Sujet de la thèse :

Traitement du signal et images LC/MS pour la recherche de biomarqueurs.

Signal processing of LC/MS images for biomarker discovery.

Thèse présentée et soutenue à Cachan le 26 juin 2009 devant le jury composé de :

Jean-Michel MOREL	Professeur des universités - ENS Cachan	Président du jury
Pierre GRANGEAT	Directeur de recherche - CEA	Rapporteur
Christine GRAFFIGNE	Professeur des universités - U. Paris V	Rapporteur
Bernard CHALMOND	Professeur des universités - U. Cergy-Pontoise	Directeur de thèse
Benno SCHWIKOWSKI	Directeur de recherche - Institut Pasteur	Directeur de thèse

Centre de Mathématiques et de Leurs Applications
(ENS CACHAN / CNRS / UMR 8536)
61, avenue du Président Wilson, 94235 CACHAN CEDEX (France)

Unité de Biologie Systémique
(Institut Pasteur / CNRS / URA 2171)
Bâtiment Laveran, 28, rue du Docteur Roux, 75024 PARIS CEDEX (France)

À Fanny
de tout mon coeur

Introduction

Présentation générale

Cette thèse de mathématiques appliquées est le résultat d'une rencontre entre un laboratoire de biologie (Unité de Biologie Systémique, Institut Pasteur) avec des problèmes de traitement du signal et un laboratoire de mathématiques (Centre de Mathématiques et de Leurs Applications, ENS Cachan) intéressé par les applications en biologie. À cette interface réside le spectromètre de masse, un instrument finalement d'une grande simplicité, qui a trouvé depuis son invention à la fin du 19e siècle des applications dans presque tous les domaines scientifiques et industriels.

L'Institut Pasteur dispose d'une plateforme de spectrométrie de masse qui réalise des analyses d'échantillons pour l'ensemble des laboratoires de l'Institut. Il s'agit principalement de caractériser la composition d'un échantillon biologique, c'est-à-dire d'identifier toutes les protéines de l'échantillon dans un premier temps et éventuellement de mesurer les concentrations de chaque espèce protéique. Les laboratoires demandent de telles analyses afin d'identifier des molécules qu'ils ont purifiées ou bien pour comparer les protéines produites par différents organismes, sous différentes conditions, etc.

L'Unité de Biologie Systémique de l'Institut Pasteur s'intéresse à l'étude du réseau formé par les interactions entre protéines d'un organisme vivant. L'approche "systémique" consiste à étudier les propriétés globales de ce réseau, sans se focaliser sur un élément particulier. Les méthodes de spectrométrie de masse fournissent de nombreuses données pour ce type d'analyses.

La mise en oeuvre d'approches systémiques en biologie moléculaire nécessite de bien comprendre le processus d'acquisition des données et les différents traitements logiciels qui sont appliqués au signal. Au travers de la collaboration entre le CMLA et l'Unité de Biologie Systémique, j'ai abordé de nombreux domaines des sciences du vivant en plus des mathématiques. Ce chapitre a pour vocation de présenter au néophyte (mathématicien ou autre) le contexte scientifique dans lequel sont effectuées les expériences de spectrométrie de masse pour la protéomique.

Un peu de biologie (systémique)

La biologie systémique étudie les propriétés d'un graphe construit à partir d'une liste exhaustive d'objets biologiques, et de la liste de leurs relations ou interactions. Différents types d'objets peuvent ainsi être étudiés :

- les séquences d'ADN (gènes)
- les protéines
- les cellules
- les individus

L'idéal de la biologie systémique est de décrire les relations entre les éléments étudiés, mais aussi les relations entre les différents ensembles. Pour cela, elle se base sur des technologies récentes

qui permettent de recenser les objets biologiques présents dans un échantillon et d'élucider leurs interactions (séquençage ADN, microarrays, spectrométrie de masse, ...).

Dans le cadre de cette thèse, nous avons étudié une technologie prometteuse — la spectrométrie de masse — pour l'analyse de l'ensemble des protéines d'un échantillon, aussi appelée le protéome. Nous présentons ici quelques éléments permettant de situer l'importance des protéines en biologie moléculaire.

ADN et Génétique

L'ADN (acide désoxyribonucléique) est une molécule qui porte l'information génétique de l'être vivant. Il s'agit des plans de construction de toutes les molécules utilisées par la cellule pour vivre. Cette information est codée par une séquence combinant quatre lettres (A, C, T et G). Dans cette séquence, on distingue un certain nombre de sous-séquences appelées gènes à partir desquelles la cellule construit des protéines.

Grâce aux nouvelles technologies de séquençage, on a déterminé le nombre de gènes présents dans le code génétique de divers êtres vivants. La principale conclusion de ces études est que la complexité des êtres vivants n'est que partiellement expliquée par la complexité de leur génome, et en particulier leur nombre de gènes. Ainsi l'être humain dispose d'environ 27000 gènes, pas plus que d'autres espèces animales (souris, 29000 gènes, oursin 23300 gènes) ou végétales (riz, 50000 gènes) [BWB03].

A l'heure actuelle, on explique la complexité additionnelle par :
- certains mécanismes de traduction du code génétique tel que l'épissage alternatif [MCS05],
- d'autres formes de code génétique comme l'ARN mitochondrial (37 gènes chez l'homme, dont 13 protéines [DMW04]),
- mais surtout les interactions et régulations dans l'expression du code génétique.

Les êtres vivants sont des systèmes dynamiques complexes, et l'on peut difficilement comprendre leur fonctionnement à partir des seuls plans statiques fournis par le code génétique.

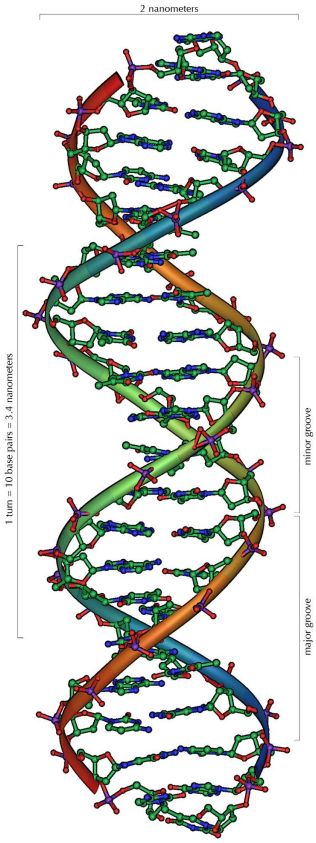

FIG. 1 – La molécule d'ADN porte le code génétique.

Les technologies actuelles ne permettent pas d'aborder l'aspect dynamique en détail, mais les approches de biologie systémique étudient certains types d'interaction entre les gènes. Elles nécessitent des méthodes exhaustives pour construire l'ensemble des gènes ou génome. On utilise aujourd'hui des méthodes de séquençage à haut débit associées à l'identification des gènes par méthodes bioinformatiques. Elles nécessitent également d'identifier les interactions entre gènes, ce qui est réalisé à l'aide de microarrays.

Un objectif futur de la biologie systémique consiste à tenir compte des interactions entre les gènes et leurs produits (ARN et protéines). En effet, la cellule produit les protéines indiquées par le code génétique. Cependant, la vitesse de production, la quantité et le type de protéines produites dépend de l'état de la cellule et de son environnement. En particulier, la présence de certaines protéines permet de réguler l'activité des gènes.

Protéomique

Plutôt que d'étudier la biologie d'un être vivant à partir du code génétique, il est parfois plus judicieux d'étudier le protéome. En effet, ce sont les protéines qui réalisent la majorité des fonctions cellulaires : métabolisme des nutriments, conversion de l'énergie, etc. En s'intéressant aux protéines, on espère une relation plus directe entre le stimulus et ses conséquences.

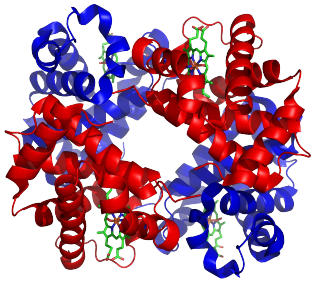

FIG. 2 – Une protéine humaine : l'hémoglobine.

Les protéines sont des molécules chimiques plus difficiles à manipuler que l'ADN. Elles ont des propriétés chimiques et des structures plus diversifiées. D'autre part, elles sont moins stables, et sont actives chimiquement ; la plupart ont une activité enzymatique.

Les protéines sont des chaînes d'acides aminés pris parmi 20. Elles ont au départ une structure linéaire. Elles sont ensuite assemblées selon plusieurs étapes de maturation qui peuvent modifier leur séquence, leur donner une structure tridimensionnelle, les regrouper en complexes. Ainsi, la molécule d'hémoglobine présentée sur la figure 2 est un complexe de quatre protéines, deux unités α et deux unités β. Les groupement hème indiqués en vert sont des modifications des protéines qui ajoutent un atome de fer et permettent le transport de l'oxygène.

La structure tridimensionnelle est un élément essentiel de la fonction de la protéine. Ainsi, la cellule peut réguler l'activité de ses protéines en modifiant leur structure, par exemple en leur greffant certains groupes chimiques qui perturbent le repliage. D'autre part, certaines maladies sont directement le résultat de dysfonctionnement dans la structure des protéines (hémoglobine falciforme, prion, ...).

La spectrométrie de masse permet d'étudier le protéome dans son ensemble et à grande échelle. Cette technique permet à la fois d'identifier la majorité des protéines présentes dans un échantillon et de mesurer leur concentration. Certaines modifications du protocole expérimental permettent d'étudier le repliement des protéines, la formation des complexes protéiques, leurs interactions, etc.

Biomarqueurs

Dans cette thèse, nous nous intéressons au protéome du point de vue de la recherche de biomarqueurs. En permettant d'identifier et de quantifier les protéines d'un échantillon, la spectrométrie de masse permet de reconnaître les différences entre individus. Un biomarqueur est alors une protéine dont la présence ou l'absence permet de caractériser l'état de l'individu. Les procédures décrites dans cette thèse sont utilisées dans certains centres hospitaliers (à titre expérimental) pour le diagnostic et le suivi clinique du cancer.

La spectrométrie de masse permet de tels diagnostics lorsque les biomarqueurs sont connus. Dans le cas contraire, la spectrométrie de masse permet aussi d'identifier des biomarqueurs potentiels, et de suggérer des candidats pour de plus amples validations.

La recherche de biomarqueurs se base sur l'hypothèse suivante. Puisque les protéines effectuent l'essentiel des fonctions cellulaires, les dérèglements associés à une maladie se manifestent probablement par des modifications du protéome : défaut dans le repliement des protéines, accumulation ou absence de certaines protéines.

Pour la recherche de biomarqueurs, la spectrométrie de masse présente deux grands avantages. D'une part, il s'agit d'une technologie sensible, capable de reconnaître des biomarqueurs dans des échantillons complexes. En particulier, on cherche souvent les biomarqueurs dans le sérum humain, car son prélèvement est à la fois aisé et sans risque pour le patient, et que l'on sait qu'il contient des traces de l'activité de l'organisme.

D'autre part, les analyses de spectrométrie de masse permettent d'identifier un grand nombre de molécules en parallèle. On peut donc soit combiner les informations de plusieurs biomarqueurs pour améliorer la fiabilité du diagnostic, soit tester plusieurs biomarqueurs dans la même expérience.

Chaîne d'analyse et traitement du signal

Les meilleurs spectromètres de masse ne sont pas assez performants pour analyser un échantillon de sérum humain sans un certain nombre d'étapes de préparation. Les analyses actuelles combinent des innovations techniques développées ces vingt dernières années qui ont permis de passer de l'analyse de petites molécules purifiées à l'analyse de mélanges complexes de protéines.

Les innovations techniques réalisées nécessitent des algorithmes de traitement du signal spécifiques, adaptés à la fois au problème expérimental et au type de données rencontrées. Le volume de données généré sur une plateforme de spectrométrie de masse, et en particulier pour les analyses à grande échelle, requiert des algorithmes performants et efficaces. Pour la recherche de biomarqueurs, il est important de se placer en limite de bruit, et d'exploiter au maximum les possibilités de l'instrument.

Dans cette thèse, nous avons abordé les procédures de traitement du signal consécutives au couplage de la spectrométrie de masse (MS) à une séparation chromatographique en phase liquide (LC). La figure 3 présente les différentes étapes d'une analyse LC/MS typique et nous suivrons cet ordre dans le manuscrit.

Nous présenterons tout d'abord la plateforme d'acquisition des mesures. La préparation des expériences et la séparation chromatographique appliquent des transformations chimiques aux échantillons et seront présentées ensembles dans le chapitre 2. Le spectromètre de masse utilise des principes de physique des particules pour la mesure de leur masse et de leur concentration. Nous en présenterons les principes et leurs conséquences pour le traitement du signal dans le chapitre 3.

Avant de présenter nos travaux sur le traitement du signal, nous présenterons l'état de l'art pour les applications de la LC/MS en protéomique. Les opérations de base consistent à identifier et à quantifier les protéines d'un échantillon et sont décrites dans le chapitre 4. On combine ces opérations pour la recherche de biomarqueurs dans le chapitre 5. Cette présentation est nécessaire pour apprécier l'impact des approches de traitement du signal développées par la suite.

Dans le chapitre 6, nous décrivons les données obtenues sur une plateforme LC/MS, et les algorithmes de calibration utilisés pour leur mise en forme. Nous nous sommes particulièrement intéressés à la calibration de la séparation chromatographique et à la calibration des intensités mesurées.

Les chapitres 7 et 8 s'intéressent au bruit dans les données. Dans le chapitre 7, nous modélisons le processus de bruit et nous décrivons une méthode d'estimation des paramètres du modèle et ses propriétés statistiques. Le modèle proposé répond aux deux principaux problèmes de calibration des intensités soulevés dans le chapitre 6. Dans le chapitre 8, on se base sur un modèle

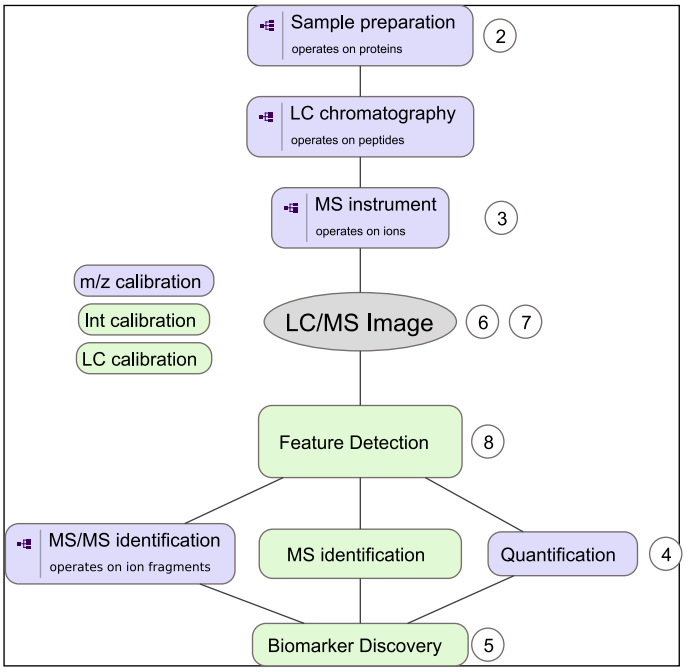

FIG. 3 – Chaîne d'analyse en LC/MS. Dans cette thèse, nous avons abordé les éléments indiqués en vert. Les chapitres de la thèse sont représentés par des numéros.

de bruit simplifié pour la détection des signaux protéiques dans les données. L'objectif est de contrôler le taux de faux positifs pour l'identification des protéines.

Les recherches présentées dans cette thèse ont fait l'objet des articles suivants :

- M. Vandenbogaert, S. Li-Thiao-Té, H. M. Kaltenbach, R. Zhang, T. Aittokallio, and B. Schwikowski. Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics*, 8 :650–672, Feb 2008.
- S. Li-Thiao-Té. Semiparametric estimation of the gain parameter with quantization errors. En préparation.
- S. Li-Thiao-Té. Feature detection with the m-n rule in liquid chromatography-mass spectrometry images. Soumis.

Le premier article a été scindé en trois parties reprises dans les sections 6.4, 4.4 et 5.2. Il aborde en effet une approche transversale restreinte de la chaîne d'analyse. Les deux autres articles sont repris aux chapitres 7 et 8.

Contents

Chapter 1

Introduction

1.1 Presentation

This thesis results from the interactions between a computational biology lab with signal processing problems (the Systems Biology Unit at Institut Pasteur) and an applied mathematics lab (the Centre de Mathématiques et de Leurs Applications, ENS Cachan) with keen interest in real-world applications. At the interface between these two scientific domains resides a simple instrument, the mass spectrometer, with widespread usage in all branches of science and industry.

The mass spectrometry platform of the Pasteur Institute is set up to analyze the various biological samples generated in the labs of the Institute. It conducts analyses with exhaustive identification of the molecules in the sample as the main objective. The platform may also conduct quantitative analyses that measure the concentration of the identified molecules in the sample. The results inform researchers on the proteins that are present in the submitted sample, differences between samples acquired in different conditions, in different living organisms, etc.

The Systems Biology Unit at Institut Pasteur is interested in the network of interactions between the proteins of a living organism. The "systems" approach consists in studying global properties of this network, without a priori focus on specific sub-networks. The data used in many "systems" approaches is generated using a mass spectrometer.

Extensive knowledge of the acquisition process of the data is necessary to fully understand and implement systems biology approaches. In particular, several preprocessing steps — both experimental and computational — may bias downstream interpretation of the results. Thanks to the collaboration between the Systems Biology Unit and the CMLA, I have been introduced to many aspects of the life sciences and of applied mathematics. The present chapter is an introduction to the context in which proteomics analyses with mass spectrometry are conducted for systems biology.

1.2 Some "systems" biology

In systems biology, the object of study is a graph that is constructed from an exhaustive list of biological objects, and from relations (*interactions*) between them. Several types of biological objects are currently investigated:

- DNA sequences (genes)
- proteins
- cells and organs
- individuals

The goal of systems biology is to explain biological phenomena based on the relationships between objects of the same type, but also relationships between objects of different types. To that end, it uses extensive data collection procedures which are enabled by recent technological advances like high-throughput gene sequencing, microarrays, 2D-gel/MS analyses, etc.

In the course of this PhD thesis, we have studied a promising technology — mass spectrometry — for proteomic applications, that is to say the exhaustive characterization of the protein content in a sample. We hereby provide some elements to outline the biological relevance of proteins in molecular biology.

1.2.1 Genetics

DNA (desoxyribonucleic acid) is the bearer of genetic information in living organisms. It provides construction plans for all the molecules involved in biological processes. Genetic information is coded with a sequence of four letters (A, C, T and G). Genes are sequences in the DNA code from which the cell assembles the proteins.

Thanks to the recent advances in DNA sequencing, the number of genes is known for a few species. The main conclusion is that the complexity of life is poorly explained by the complexity of genetic information, as expressed in terms of the number of genes in the DNA sequence. For example, the human genome contains roughly 27,000 genes, which is no more than other animal species (mouse, 29,000 genes, sea urchin 23,300 genes) or plant species (rice, 50,000 genes) [BWB03].

Additional complexity is currently explained by

- mechanisms in gene translation like alternative splicing which produce several different proteins from the same DNA sequence [MCS05],
- other forms of genetic code like mitochondrial RNA (37 genes in human cells, among which 13 are proteins [DMW04]),
- and in particular interactions and regulations in the expression of the genetic code.

Living organisms are complex dynamical systems, and can hardly be understood from the static plans provided by the DNA sequence.

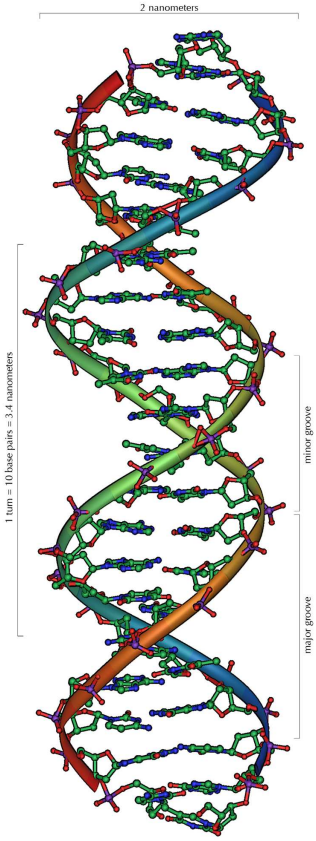

Figure 1.1: Desoxyribonucleic Acid.

Current technology cannot tackle the dynamics of biological systems in details. However, systems biology approaches can probe some aspects of the interactions between genes. They make use of high throughput sequencing technology to obtain the set of genes. Microarrays, yeast-two-hybrid and other methods can subsequently be used to determine interactions between genomic sequences and consequently genomic interactions.

A future objective of systems biology is to take into account interactions between genes and their products (RNA and proteins). For example, cells produce proteins based on the DNA sequence in varying rates and quantities, depending on its needs and the environmental conditions in its surroundings. In particular, some proteins are signals to increase or lower the activity of genes.

1.2.2 Proteomics

The biological processes of a living organism can be studied based on its proteins rather than based on its genetic information. Most of the cellular functions are carried out by proteins: energy metabolism, nutrient intake and conversion, signalling, transport, etc. Proteins are thus generally better related to a given stimulus than the immutable genetic information in the genome.

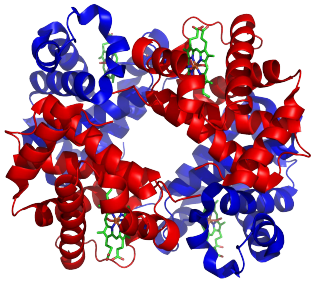

Figure 1.2: Hemoglobin, a human protein.

As chemical compounds, proteins are more difficult to handle than DNA molecules. Proteins have more diverse chemical properties and structure. They are less stable and are usually chemically active; most have an enzymatic activity.

Proteins are chains of amino-acids chosen among 20 variants. When translated from genetic information by the ribosome, they have a linear structure. They subsequently undergo maturation phases where their sequence may be altered, they adopt a three-dimensional conformation or may be linked to other proteins to make complexes. For instance, the hemoglobin molecule depicted in Figure 1.2 is a protein complex made of four proteins: two α units and two β units. Modifications to the amino-acid chain indicated in green are called heme groups; these control the fixation and release of oxygen.

The three-dimensional structure of a protein is essential to its function. Indeed, one of the processes by which cells regulate the activity of proteins consists in modifying their structure, by attaching chemical groups that affect the stability of protein folding. Some diseases are caused by mis-folding of proteins. In the case of amyloidoses (prion diseases, Alzheimer's disease, Parkinson's disease), excessive quantities of mis-folded proteins accumulate in the cells. Some types of cancer may be caused by mis-folded p53 proteins, which lose the ability to repair DNA damage.

Mass spectrometry is a high-throughput technology to analyze the proteome. The methodology presented in this thesis can be used to identify large sets of proteins in a given biological sample as well as measure their concentration. Modifications to the protocol can be used to study protein folding, the formation and dynamics of protein complexes, protein interactions, etc.

1.3 Biomarkers

Large-scale identification and quantification of the protein content of a biological sample is of particular importance to biomarker discovery. Biomarkers are proteins which can be used to characterize the biological state of an individual, diagnose diseases or monitor their progress. Some clinical studies are already using mass spectrometry for early diagnosis of cancer, and also to determine the cancer subtype.

Mass spectrometry can be used to detect known biomarkers in biological samples. In this thesis, we are rather interested in finding new and potential biomarkers based on mass spectrometry data. After validation, these potential biomarkers can be used to detect diseases but can also indicate potential drug targets for the pharmaceutical industry.

As proteins are deeply involved in all cellular processes, diseases are reflected in the proteome by mis-folding of proteins, accumulations of proteins, or absence of some proteins. Biomarkers are typically sought for in human serum, a highly complex blood extract. Serum is easy to collect and the procedure incurs no risk for the patient, unlike invasive procedures like biopsy. It is also

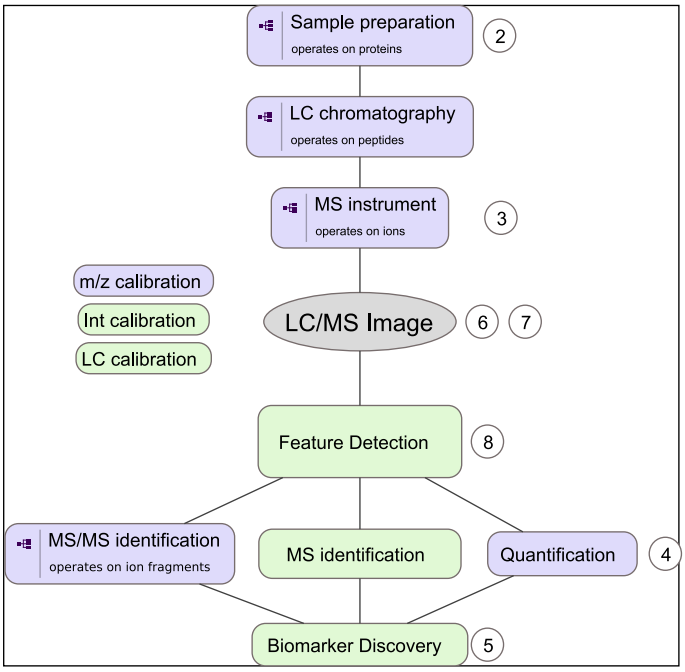

Figure 1.3: Typical LC/MS analysis pipeline. Our work tackles the areas highlighted in light green. Circled numbers indicate the chapters in which the topic is discussed.

believed to contain trace amount of the proteins produced by the cells throughout the body, and thus reflect the state of all organs.

For biomarker discovery, mass spectrometry has two main advantages. It can detect very low amounts of proteins and it can provide information on a large set of proteins in a single experiment. Consequently, biomarkers can be detected in parallel. It is also possible to combine weak information provided by several biomarkers to increase the confidence in the diagnostic.

1.4 Analysis pipeline and signal processing

Several preparation steps are required before mass spectrometry analysis of a complex sample like human serum. Mass spectrometry analyses have previously been focused on small molecules, but recent technological developments now enable the study of larger molecules like proteins.

Specific signal processing algorithms have been developed to cope with the new instrumentation and protocols and the results of their application to complex protein mixtures rather than purified proteins. Mass spectrometry platforms generate vast amounts of data, especially during large-scale experiments. This creates the need for quick and efficient algorithms. For biomarker discovery, sensitive algorithms are needed to detect the potential biomarkers as these are anticipated to correspond to low intensity signals.

In this thesis, we have studied signal processing for data generated on mass spectrometry platforms (MS) that are coupled to liquid-phase chromatographic separations (LC). Figure 1.3 provides an overview of the different steps from sample preparation to biomarker discovery.

We start with the LC/MS platform. Chemical preparation of the sample is carried out during sample preparation and liquid chromatography separation. Both will be presented in Chapter 2. Mass spectrometry analysis uses particle physics to separate proteins. The different instruments and their principles of operation are presented in Chapter 3. In both chapters, we elaborate in particular on the elements that are relevant for identification, quantification and biomarker discovery.

To motivate our signal processing methods, we present the state of the art for proteomics applications of LC/MS instrumentation. The basic tasks are to identify all the proteins in a sample and to measure their concentration. This is explained in Chapter 4. Identification and quantification results are combined and compared between experiments for biomarker discovery as presented in Chapter 5.

In Chapter 6, we present the data acquired on an LC/MS platform, and the preprocessing steps for data clean up. Preprocessing is of particular importance, and strongly affects the subsequent analyses. We have studied in more detail the calibration of the chromatographic separation and of recorded intensity values.

Chapters 7 and 8 deal with background noise in the data. In Chapter 7, we propose a model for the background noise and its statistical properties. The model provides a new method to calibrate intensity values. In Chapter 8, we study statistical signal detection in the data. The objective is to control the false positive rate for identification of the protein signals in the data.

Research presented in this thesis has lead to the following manuscripts:

- M. Vandenbogaert, S. Li-Thiao-Té, H. M. Kaltenbach, R. Zhang, T. Aittokallio, B. Schwikowski. Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics*, 8:650–672, Feb 2008.
- S. Li-Thiao-Té. Semiparametric estimation of the gain parameter with quantization errors. In preparation.
- S. Li-Thiao-Té. Feature detection with the m-n rule in liquid chromatography-mass spectrometry images. Submitted.

To facilitate the organization of the document, the first manuscript has been split and is reproduced in Sections 6.4, 4.4 and 5.2 because it tackles restricted aspects of the analysis pipeline. The two other manuscripts make most of Chapters 7 and 8.

Notations and symbols

Notations

LC	Liquid Chromatography
MS	Mass Spectrometry
Da	Dalton, mass unit synonymous to the Atomic Mass Unit (a.m.u). It is defined as a twelfth of the mass of a ^{12}C carbon atom in its ground state. The mass of a proton and a neutron are roughly equal to 1 Da.
m, m/z	mass-to-charge ratio of an ion. z is the absolute value of the charge.
t, rt	retention time
$\mathcal{I}$	intensity measured by the mass spectrometer
$\mathcal{S}$	peptide signal intensity
$\mathcal{N}$	background noise intensity
A	area under the curve
Γ	Gaussian function
σ	standard deviation of a Gaussian function

Glossary

retention time	instant at which a protein emerges from the chromatography column. It is related to the protein's chemical properties.
fraction	simplified sample obtained by purification, extraction, etc.
isotope	same molecule with a different number of neutrons
peptide	small protein with less than one hundred amino-acid residues
identification	determination of the sequence of a protein
quantification	measure of the concentration of a protein
biomarker	protein with explanatory power in terms of disease classification

Chapter 2

Liquid chromatography

Figure 2.1: Gradient liquid chromatography apparatus.

Biological samples considered in proteomics studies are often too complex for analysis based on a single mass spectrum. Consequently, current methods use liquid chromatography to fractionate the sample, and mass spectrometry is applied to each of the fractions.

In the current chapter, we present the sample preparation steps that precede mass spectrometry itself. We deal successively with chemical sample preparation of the biological sample, enzymatic digestion of the proteins with trypsin, and chromatography separation. In particular, we explain some of the compromises that are made during sample preparation and how these affect the signal measured in the mass spectrometer. For additional information, we refer the reader to [Lan05, McM07, SKG97].

2.1 Sample preparation

2.1.1 Inactivation of the proteins

Proteins found in living organisms often have an enzymatic activity and can increase the speed of chemical reactions[1]. In particular, the cellular function of some proteins is to degrade other proteins. The action of those proteins must be prevented for storage of the biological sample, or simply to prevent its transformation during the course of the experiment.

[1]Some proteins do not have an enzymatic activity. Some are used for their structural properties, e.g. keratin which is the main component of nails and hair or collagen that gives structure to most tissues in mammals. Other proteins are used as molecular signals or hormones.

Sample integrity has to be preserved in the short term because experimental analysis of a sample may take days or weeks for some protocols. This is the case for analyses which use extensive fractionation of the biological sample, or simply when the analysis is repeated on a sample to evaluate the reproducibility of the measures. Some samples are kept in the long term, for example during clinical trials. This allows the comparison of samples taken at different time points, with randomization of the experiments.

Sample preparation in LC/MS is the result of several contradictory objectives:

- inactivate the proteins to preserve the state of the biological sample,
- ensure compatibility with liquid chromatography separation techniques,
- ensure compatibility with protein ionization for mass spectrometry analysis.

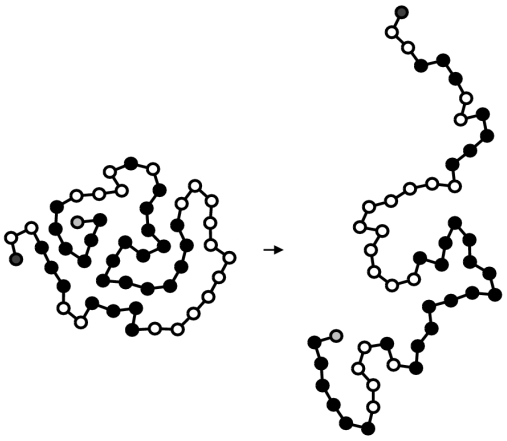

Figure 2.2: Denaturation alters the folding structure of a protein.

Proteins are inactivated by introducing denaturing chemical compounds in the sample. The first stage is usually to undo the three-dimensional structure of the proteins, and make them adopt an unfolded linear structure. As some proteins fold naturally, alkylating compounds are introduced to prevent the formation of secondary structure and chemical bonds between the amino acids.

In the long term, biological samples are stored frozen as cold reduces protein activity, and dessicated by vacuum centrifugation for example. The sample is re-solubilized before liquid chromatography separation, but not necessarily in water. Depending on the actual protocol, the proteins can be better analyzed in organic solvents like methanol or formic acid.

Chemical compounds added in sample preparation are undesired contaminants. They are sometimes visible the data set and they interact with the mass spectrometry measurements. Although most of these are small molecules, they can aggregate and form large complexes with size comparable with the proteins.

The compounds can also saturate the mass spectrometer and hide the protein signal. Consequently, sample preparation protocols are often optimized for some known proteins that are expected in the results, at the risk of missing others.

Finally, some chemical compounds can change the elution order of the proteins in the separation. This prevents the correction of retention time differences between LC/MS images as described in Section 6.2.

Conclusion Protein denaturation is a required step in the preparation of biological samples for LC/MS. Unfortunately this modifies the properties of the solvent and of the sample, which makes it difficult to compare the data collected from different labs.

2.1.2 Protein extraction / depletion

There are several reasons to extract part of the protein population in a sample:

1. Proteins like albumin and immuno-globulins make the majority of proteins in human serum. These highly abundant proteins are usually filtered out (depletion) to enhance signals from other proteins.
2. When dealing with cells, the membranes have to be broken, and protein extracted from the biological matrix.
3. Proteins have a wide range of chemical properties. In particular, membrane proteins are difficult to dissolve in aqueous solvents. Membrane proteins and cytosomal proteins can be analyzed separately using adequate protocols.
4. Protein extraction can focus on a restricted population of proteins. For example, centrifugation can be used to extract only the proteins that correspond to a specific organelle in the cell (ribosomal proteins, etc.)

The desired proteins are extracted by separating the components of the biological sample. Here are some examples of the techniques involved:

- Precipitation. Depending on temperature, pH or solvent composition (addition of ammonium sulfate for example), some proteins precipitate, i.e. form solid compounds that can be separated from the liquid phase.
- Immunoprecipitation / affinity capture. Some antibodies bind specifically to proteins. They can be bought from manufacturers and immobilized on metal beads or glass plates. After the proteins are bound to the antibodies, they can be pulled from the biological sample, and separated from the solid support.
- Centrifugation can be used to separate blood into serum, red cells and platelets. It can also be applied to separate molecules based on their size.
- Chromatography is described in Section 2.3.

2.2 Enzymatic digestion

After denaturation, proteins are no longer chemically active. However, they still have a wide range of chemical properties, and it is difficult to analyze all of them in one experiment.

Peptides are small proteins — only up to a hundred or so amino acids — that have more homogeneous chemical properties. Current LC/MS analyses operate on peptides because a single protocol is sufficient and that protocol can be automated. Moreover the mass spectrometer can focus on a smaller mass range so that the obtained mass spectra are more detailed and the acquisition time is reduced.

Proteins in a biological sample are transformed into peptides by an enzyme which cuts proteins into pieces. Trypsin is the most commonly used protease in LC/MS-based proteomics. It selectively cleaves the protein sequence after lysine (K) and arginine (R) residues. Tryptic digestion is set up so that the peptide fragments obtained from the proteins are minimal and cannot be cut

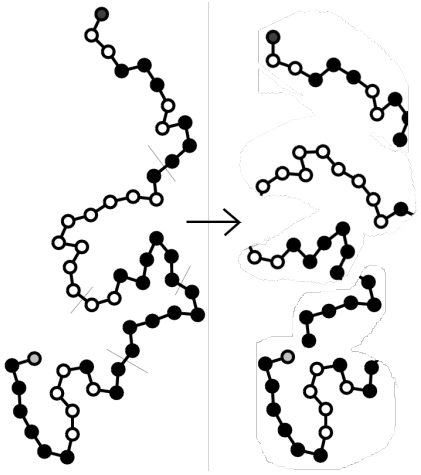

Figure 2.3: Cutting a protein into peptides by enzymatic digestion.

further. It has the disadvantage that the sample contains many more molecular species; some proteins can be broken into dozens of fragments.

Example: Tryptic digest of insulin
One form of insulin has the following amino acid sequence:

```
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFY\
TPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

After tryptic digestion, the following peptides are obtained:

```
MALWMR
LLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGER
GFFYTPK
TR
R
EAEDLQVGQVELGGGPGAGSLQPLALEGSLQK
R
GIVEQCCTSICSLYQLENYCN
```

After identification of the peptides in the sample by mass spectrometry, the results are assembled to identify proteins. Each peptide is associated to its corresponding protein; a protein is considered as identified in the sample depending on the number of identified peptides, and three present peptides are usually considered conclusive evidence.

A given peptide can be uniquely attributed to a protein as soon as it is long enough. With 6 or 7 amino acid residues, the number of combinations of the 20 possible amino acids is around 64 million, and greatly outnumbers the number of human genes (25,000) or human proteins (100,000).

From a peptide sequence, the corresponding protein can be found in a protein database. Exhaustive protein libraries can be constructed based on genomic information, and large-scale sequencing technologies are available today. Proteins are usually long enough so that several of its peptides can be used for identification. The main difficulty is to integrate the measurements obtained from several peptides into a coherent measure for the protein.

Remark Trypsin needs to be inactivated after digestion. Otherwise, even though it is specific to lysine and arginine residues, it can cut at other sites and produce unexpected non-tryptic peptides. In past versions, trypsin could cut itself and produce trypsin autolysis fragments in the list of identified peptides. Current commercial versions are modified to prevent autolysis. Additionally, if the chemical reaction is incomplete, there may be miscleaved peptides in the resulting peptide mixture, for instance `GFFYTPKTR` in the previous example.

Conclusion Nearly all LC/MS analyses for proteomics are carried out after tryptic digestion of the biological sample. This increases the range of proteins that can be analyzed with a single experiment on the LC/MS platform. However, proteins need to be reconstructed from their peptide fragments.

In this thesis, we will make no difference between proteins and peptides, and we will use both terms interchangeably. We assume that peptides can be uniquely attributed to their corresponding protein. Protein identification is performed on the basis of identification of peptides. Likewise protein quantification is deduced directly from the quantification of peptides. How to properly combine the measurements from several peptides of the same protein is still an open question.

2.3 Liquid chromatography separation

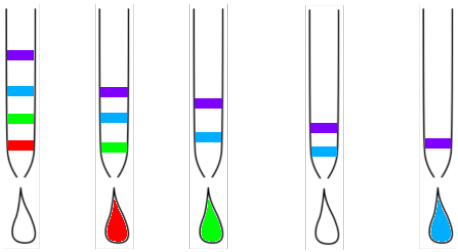

Figure 2.4: Liquid chromatography with a column. As time passes, proteins represented by colored bands progress down the chromatography column. The instant at which a peptide leaves the column is called its retention time.

Liquid chromatography is a technique to fractionate a mixture of proteins. It is widely used in conjunction with mass spectrometry because it is easy to interface with mass spectrometry and acquire mass spectra for each fraction.

Liquid chromatography improves mass spectrometry analyses because each fraction is analyzed independently. Fractions introduced in the mass spectrometer are simpler and each protein has a higher relative concentration.

2.3.1 Separation principle

A liquid chromatography column is packed with porous material. When the sample is introduced in the column, the proteins interact with the packing material and their progression is slowed down depending on their chemical properties. The *retention time* of a protein corresponds to the instant it leaves the chromatography column.

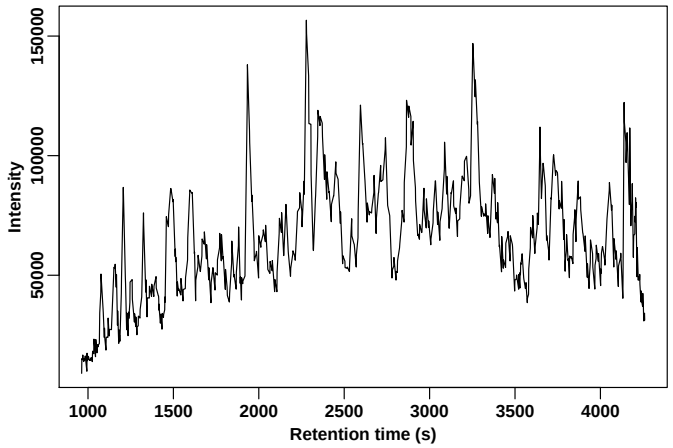

Figure 2.5: A Total Ion Chromatogram (TIC) represents the total amount of proteins as a function of retention time. The proteins are properly separated when there are peaks throughout the retention time range.

Proteins can be separated according to three types of chemical properties:

- ion-exchange columns separate molecules based on their charge,
- size exclusion columns separate based on the size of the molecules,
- affinity columns bind to the proteins based on other chemical properties like hydrophobicity, or affinity with the packing material.

The separation power of a liquid chromatography column is improved when the packing material is smaller. This has lead to the development of nano-LC separations which improve the sensitivity of the platform as well as its separation power. Nano-LC methods also require lower amounts of biological material. Several fractionation steps can be combined (e.g. the MudPIT approach [LAYI01]) but often at the cost of increased time of analysis and manual interventions on the instrument.

In reverse phase chromatography (RPLC), proteins are first loaded on the chromatography column. Then they are eluted (released) when solvent is flowed through the column. By manipulating the solvent composition during the course of the experiment, experimentalists obtain some level of control on the elution time and separation power of liquid chromatography.

Figure 2.1 shows a common experimental setup where the solvent composition is modified by mixing two solvents before injection in the column. This procedure, called *gradient elution*, improves the separation power by focusing each protein to a small number of fractions. Figure 2.6 displays solvent composition as a function of time when using a linear gradient. Other types of gradient have been reported, such as exponential gradients or step gradients.

2.3.2 Reproducibility problems in liquid chromatography

As the different separation techniques are based on different chemical principles, they produce different results, and in particular, the elution order of proteins is not always comparable. However, even repeated analyses of the same platform with the same experimental protocol can show large differences.

Liquid chromatography is difficult to reproduce faithfully because it is very sensitive to environmental conditions: room temperature, pressure, etc. The liquid chromatography column also

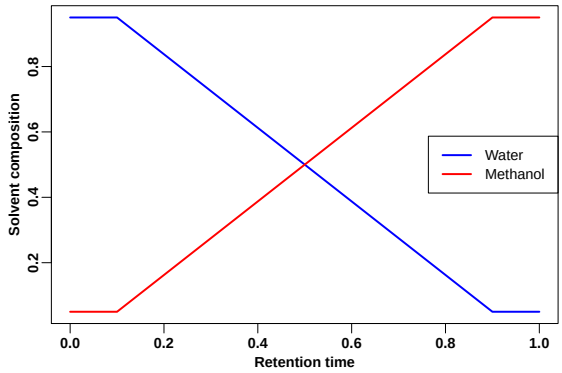

Figure 2.6: Evolution of the solvent composition in gradient RPLC.

has a limited lifespan of only 1,000 injections. Its separation performance evolves during its lifetime with the deterioration of the packing material and accumulation of chemicals that sustain extensive washing.

The sample preparation procedures presented in the previous sections also have an impact on liquid chromatography. They can modify the properties of the solvent and change the retention time of proteins. They bind to the packing material and may clog under certain conditions.

Experimental variations of the chromatography separation affect the retention times of the proteins. These are mostly global shifts of the retention times, and the relative order of elution is preserved. Local inversions of peptides are possible but rare. This property is the basis of the chromatography correction algorithms presented in Chapter 6. More details and references can be found in [VLTTK+08] (also reproduced in Chapter 6 on page 90).

Summary

In this chapter, we have presented the chemical modifications of the biological sample that precede its analysis by mass spectrometry. To process complex mixtures of proteins the following steps need to be applied:

1. inactivation of the proteins to ensure sample integrity,
2. enzymatic digestion of the proteins into peptides to increase coverage,
3. separation with liquid chromatography to improve the detection of low-abundance peptides.

These sample preparation steps create aberrations that can be corrected with signal preprocessing algorithms. Contaminants create chemical noise, which is the main component of noise in the data. Tryptic digestion implies that protein data has to be reconstructed from several measurements. Liquid chromatography is useful for peptide identification as discussed in Chapter 4, but its non-reproducibility is a major issue. Chapter 6 presents methods to correct the observed retention times.

Sample preparation as described in this chapter is applied in the context of large-scale proteome analyses with exhaustivity as the main objective. As there are no a priori on the types of protein of interest, this approach is well suited to the untargeted discovery of potential biomarkers.

When focusing on known proteins, some or all of the previous steps can be skipped. For example, when studying protein complexes, denaturation will destroy the bonds between the proteins. Tryptic digestion is not necessary and whole proteins can be analyzed on some mass spectrometers, depending on their size. Complexes are often extracted from the biological samples and liquid chromatography may not be needed to separate them.

Chapter 3

The mass spectrometer

Figure 3.1: A mass spectrometer.

3.1 Overview of mass spectrometry

A mass spectrometer *separates* a flux of *ions* according to their mass-to-charge ratio, then measures the *intensity* of the ion beam by counting the number of ions that impact a detector.

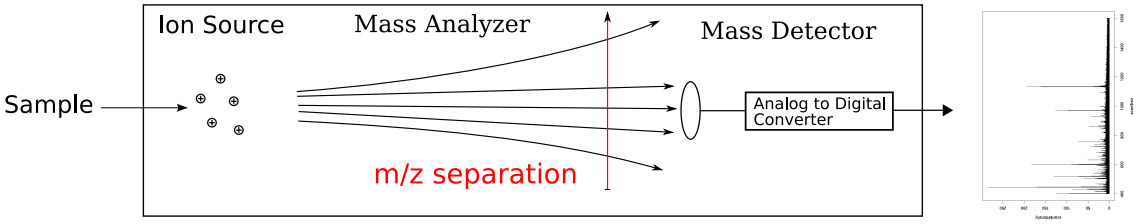

Figure 3.2: Schematic representation of a mass spectrometer

This principle of operation was discovered by Thomson at the end of the 19th century, with the observation that ion beams are bent by magnetic fields. Mass spectrometry is a technique of

choice because it is very sensitive, quantitative, and because the detected ions can be identified based on their mass-to-charge ratio.

In the present chapter, we describe the elements of a mass spectrometer:

- the ionizer, which creates ions from the proteins in the sample,
- the mass analyzer, which separates the ions based on their mass-to-charge ratio,
- the ion detector, which measures the intensity of the ion beam.

We present the technical choices for each component, and their consequence on the type of data acquired by the mass spectrometer. Although we have focused on TOF instruments in the course of this thesis, the other instruments are widespread and our research applies to most of them in spite of the technical differences. Additional information can be found in [McM07, Lan05, DA06].

3.2 Ionization

In LC/MS analyses, the proteins are dissolved in the liquid phase when they arrive at the mass spectrometer. However, mass analysis operates on gas phase ions in vacuum. The ion source has two tasks. The first one is to transfer the peptides from liquid phase to gas phase. The second task is to provide peptides with positive or negative charges.

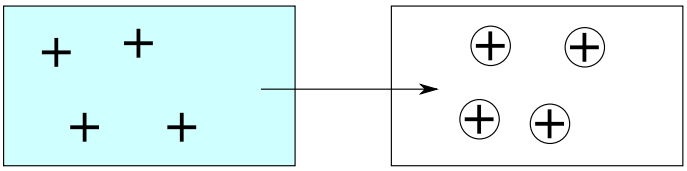

Figure 3.3: The ion source transfer the sample peptides from liquid phase (in blue) to gas phase (in white). The created ions are then transferred into the vacuum of the mass analysis chamber.

In this section, we describe the two main ionization techniques in LC/MS: electrospray ionization and MALDI ionization. The ionization technique determines what types of signals are observed in the data, and background noise is different in the two methods.

3.2.1 Consequences of ionization on the data

The ideal case is when the intensity of the ion flux produced in the ion source is proportional to the concentration of the protein in the sample. Intensity corresponding to different proteins add to the total intensity. This linear mode of operation is limited by the following experimental phenomena.

Ionization problems There are many ionization techniques besides electrospray and MALDI. These two are the main techniques in LC/MS because they are soft ionization techniques. As they operate at low energy levels, the proteins are preserved during ionization and remain intact [1].

[1]The parameters are usually set such that the proteins remain intact. It is possible to induce fragmentation in the ion source, which may be useful to replace MS/MS mode, see Section 4.2

Ionization efficiency Each ionization technique has its preferred substrates. For instance, in electrospray ionization, hydrophobic proteins are more easy to ionize than hydrophilic proteins. Ionization efficiency, i.e. the number of ions divided by the number of proteins, is very different depending on the chemical properties of the molecule (polarity, affinity, hydrophobicity, etc.) but also depends on the experimental conditions (temperature, pressure, etc.). A major issue is ionization suppression, when some of the molecules in the sample suppress the ionization of other molecules [TPS04].

Saturation The ion source has a limited capacity. When the sample protein concentration is high, only the most abundant are ionized, and low-abundance proteins are suppressed in the signal. There is no longer a linear relationship between sample concentration and measured signal.

Chemical noise The ion source creates chemical noise, but the process is not well understood. The ion source is thought to:

- let some of the ions reach the detector without separation in the mass analyzer
- create molecular complexes that randomly fragment into ions during the various stages of mass analysis.

3.2.2 Electrospray ionization (ESI)

Principle summary In electrospray ionization, the liquid sample is introduced in capillary tubing. A high voltage difference (1 to 6 kV) is applied between the tip of the capillary and the entrance of the mass analyzer chamber. Charges accumulate at the tip and form droplets that carry the molecules to the mass analyzer.
More detailed information can be found in [Gas97, CE01]

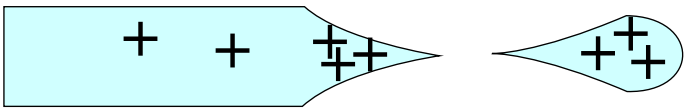

Figure 3.4: Electrospray ionization. The voltage difference applied between the capillary and the mass analysis chamber creates an accumulation of charges at the tip of the capillary and the formation of charged droplets.

Details When a positive voltage is applied to the capillary, proteins with a positive charge accumulate at the tip. This modifies the surface properties of the liquid phase because the voltage creates a force with compensates surface tension. Under the right experimental conditions, a stable *Taylor cone* is formed at the tip.

At the extremity of the Taylor cone, the accumulation of charges increases electrostatic repulsive forces between the ions in solution. When the forces exceed the surface tension of the liquid (Rayleigh limit), small droplets (around 10 μm in diameter) of sample are liberated. These droplets are charged and they are attracted to the entrance of the mass spectrometer.

Charged droplets contain both solvent molecules and proteins in liquid phase. Before entering the mass analyzer, the droplets are fragmented further. Droplet fragmentation is the result of two phenomena:

- during the flight of the droplet, the solvent evaporates, especially if heated. Because of increasing electrostatic forces in the shrinking droplet, the droplets fragment and proteins are separated.

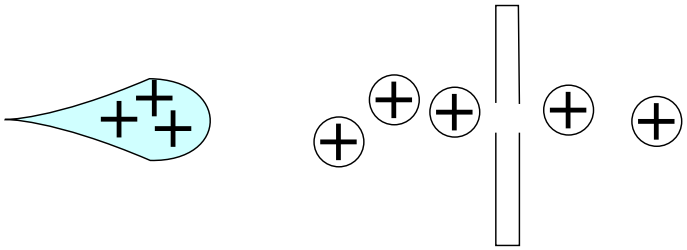

Figure 3.5: As droplets are further fragmented, they liberate the peptides and transmit their charges. The created peptides are focused on the entrance of the mass analysis chamber.

- an equivalent of the Taylor cone is formed at the rear end of the droplet. Smaller droplet are liberated in a trail.

Droplets are fragmented until the solvent molecules are separated from the proteins and each protein is by itself, i.e. in gas phase. Charges on the droplet are transmitted to the proteins it contains.

In electrospray ionization, the ionization process takes place at atmospheric pressure, in contact with the atmosphere. The capillary is outside the mass spectrometer. The entrance of the mass spectrometer is thus open to the lab atmosphere, so volatile molecules (perfume, wall paint) can enter the mass analyzer and contaminate the results.

The capillary used for electrospray ionization is directly connected with the liquid chromatography column. As chromatography cannot be interrupted, there is a constant flow of analyte molecules in electrospray. If these analytes are not analyzed somehow, they are definitively lost.

Type of signal Electrospray ionization typically creates peptide ions with two to three charges; singly charged peptides and higher charge states are less represented in the results. Proteins with hydrophobic properties are easier to ionize because they naturally accumulate at the surface of the solvent.

Positive ion mode is well compatible with tryptic peptides because lysine and arginine residues are negatively charged and attract positive charges easily. Negative ion mode can be performed under different pH conditions.

3.2.3 Matrix-Assisted Laser Desorption Ionization (MALDI)

Principle summary The sample is deposited on a metallic plate coated with matrix, and placed inside the mass spectrometer. A laser pulse excites and vaporises a thin layer of the matrix molecules and the proteins. This results in a cloud of gas phase ions above the MALDI plate.
More details can be found in [HPK07].

Details Sample is deposited on a metallic plate that has been coated with a polymer forming a matrix. Solvent from the sample is then dried, and the proteins are crystallized in the matrix. As the matrix molecules are largely in excess of the proteins, the proteins are isolated in the matrix. Matrix and proteins form a solid material on the metallic plate.

When hit by a laser pulse, a thin layer of the matrix + protein compound is vaporised, and forms a cloud of ions above the plate inside the vacuum of the mass spectrometer. This clouds contains gas phase proteins as well as clusters of matrix molecules. The matrix polymer assists the ionization of proteins by providing charges. However, complexes of matrix molecules contribute greatly to chemical noise in MALDI mass spectra.

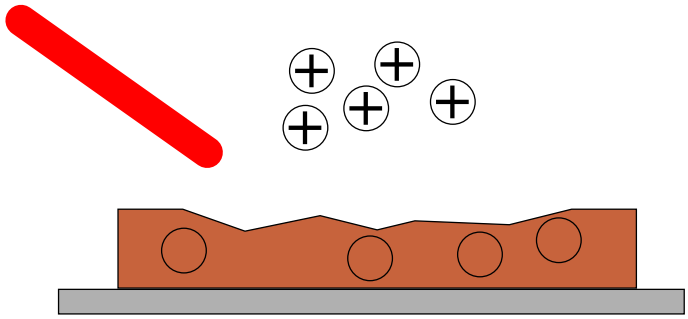

Figure 3.6: MALDI ionization. When hit by a laser pulse, a thin layer of matrix and sample are vaporised. Ions are formed in the plume abome the MALDI plate.

As the sample proteins are immobilized inside the matrix, the sample can be preserved for future analyses. With each laser pulse, only a thin outer layer is destroyed, and enough proteins remain. One of the problems in MALDI ionization is that the proteins are not uniformly distributed in the matrix. Consequently, mass analysis is conducted by accumulating several laser shots from different locations.

Type of signals Maldi ionization usually creates ions with a single positive charge.

Conclusion

Two main ionization technologies are used in LC/MS analyses for proteomics, electrospray ionization and MALDI ionization. Both can be used for large molecules like proteins, and the ionized molecules are kept intact.

Electrospray and MALDI do not produce the same signals in the mass spectrometer. One reason is that electrospray generates ions with more charges. Additionally, the ionization efficiency of proteins is different in the two technologies.

Both technologies are widely used in proteomics facilities. One advantage of MALDI ionization is that the sample can be stored for further analysis. However, electrospray is easier to interface with liquid chromatography: it is possible to consider a higher number of fractions. Another difference is that MALDI methods are usually used in high-throughput analyses, whereas electrospray methods have a better sensitivity to low concentration proteins.

3.3 Mass analysis

The mass analyzer is the central part of the mass spectrometer. The quality of the results obtained on a mass spectrometry platform is largely determined by the capacity of the mass analyzer to separate ions based on their mass-to-charge (m/z) ratio. The cost of the platform as well; a mass spectrometer can cost a million euros.

Principle Charged particles move inside electromagnetic fields according to Maxwell's equations. Mass spectrometry consists in designing the electromagnetic field so as to separate the peptide ions based on their mass-to-charge ratio. Mass analysis operates in vacuum because the presence of air molecules would slow down the movement of the peptide ions and prevent its observation.

We can distinguish two main types of mass analyzers:

- scanning mass analyzers act as a mass-to-charge ratio filter. Only the molecules with a specific m/z ratio have a stable trajectory in these analyzers (Magnetic sector, quadrupole, TOF).
- ion traps are devices where the electromagnetic field is used to confine the ions inside a trap. For detection, ions can be selectively ejected from the trap (Ion trap, FT-ICR).

In this section, we present the characteristics of a mass analyzer that are relevant to signal processing of data in LC/MS. For each type of mass analyzer, we describe its implementation, and the performance attained in practice. We will use the terms "mass" and "mass-to-charge" interchangeably.

3.3.1 Characteristics of a mass analyzer

A mass analyzer can be described by the following characteristics:

- mass accuracy and precision
- mass range
- dynamic range
- sampling rate
- scanning speed

The most important characteristics of a mass analyzer are its accuracy and precision, which define quality of the m/z measurements. Accuracy defines whether repeated measures correspond to the theoretical value (Law of Large Numbers). In statistics, this corresponds to the bias of an estimator. Precision defines whether repeated measures are close to each other (Central Limit Theorem). It corresponds to the standard deviation of an estimator.

Mass accuracy In mass spectrometry, accuracy can be defined as the difference between the mass of the centroid of a peak in the mass spectrum, and the theoretical mass of the corresponding ion, as computed from its chemical formula[2]. Mass accuracy is usually measured in ppm (parts per million), and given for a molecule of mass around 1,000 Da. For example, a mass difference of 0.1 Da corresponds to 100ppm accuracy[3].

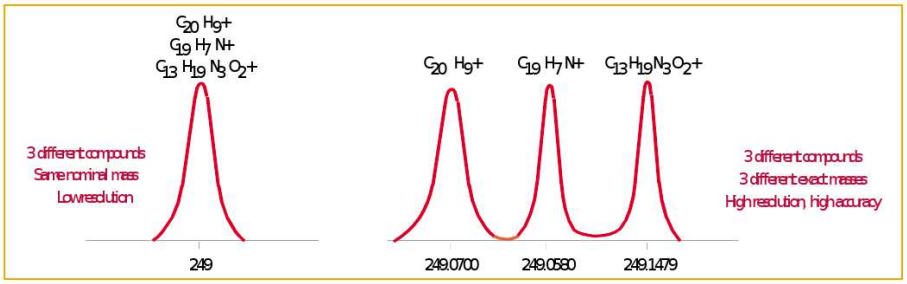

Figure 3.7: Accuracy of a mass analyzer.

Mass precision The precision or resolving power of a mass spectrometer corresponds to the capacity of distinguishing ions of close mass-to-charge ratios. The resolution can be computed according to two different methods illustrated in Figure 3.8. Resolution has no unit.

[2]See the appendix for details on this computation.
[3]One Dalton corresponds to the mass of a hydrogen atom or to the mass of a proton.

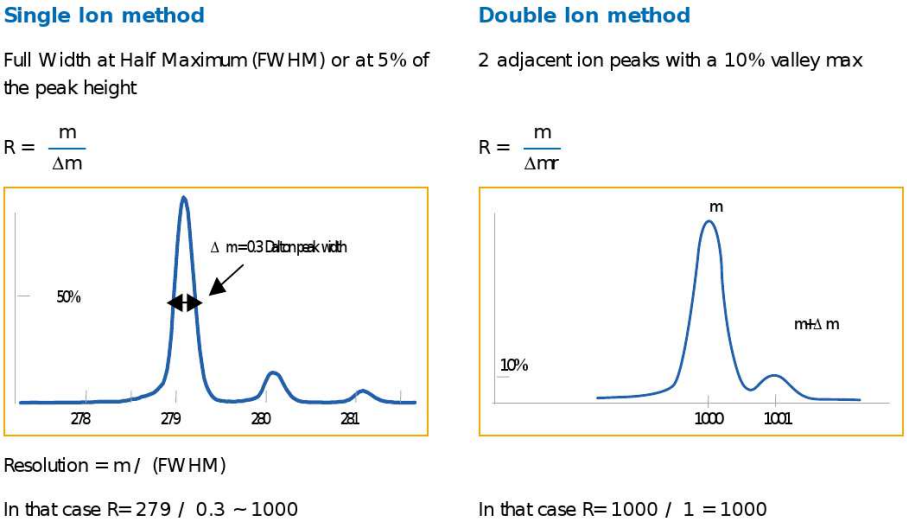

Figure 3.8: Resolution of a mass analyzer.

Mass range Mass analyzers can separate ions of low molecular weight and up to a certain limit. TOF instruments in particular can analyze very large molecules like intact proteins. All current mass spectrometers can analyze the peptide fragments obtained by tryptic digestion during sample preparation (see Section 2.2).

Dynamic range expresses the relative intensity of ions that can be analyzed without saturation of the instrument. Some mass analyzers, in particular the ion traps, have a limited dynamic range.

Sampling rate According to the type of mass analyzer, the sampling rate of the m/z axis is different. In particular, the sampling rate is not uniform for TOF instruments.

Scanning speed The scanning speed is the time required to acquire a mass spectrum. Of course high scanning speeds are preferable. They require less analysis time, and can be used in conjunction with quicker LC separations, or to provide more detailed data. The scanning speed of an instrument can be adjusted but speed is usually traded with instrument sensitivity.

3.3.2 Magnetic sector mass analyzers

Magnetic sector mass analyzers were the first available mass spectrometry systems. They are not normally used in LC/MS platforms because of the slow scanning speed. We present these instruments as an introduction to the more complicated setups.
More details can be found in [BH96].

Principle Magnetic sector analyzers use the same physical principle as cyclotrons. Ions travelling in a constant, orthogonal magnetic field adopt a circular trajectory at constant velocity. The radius of that trajectory is dependent on the m/z ratio of the particle.

Details Let m denote the molecular mass of an ion of charge z. After ionization, the ion is accelerated in an electric field up to an energy $E = zV$ where V denotes the potential difference. The particle moves in a straight line at speed v such that $E = \frac{mv^2}{2}$.

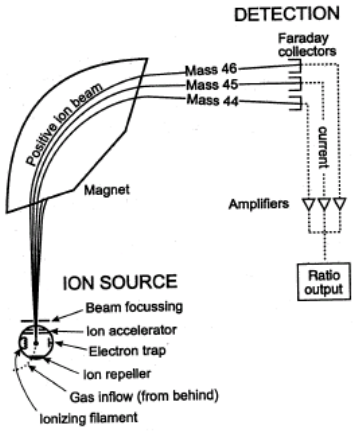

Figure 3.9: Magnetic sector mass analyzer.

The ion is then introduced in the magnetic sector. In that portion of the mass spectrometer, the ion trajectory is bent by the presence of a high intensity constant orthogonal magnetic field. The radius of the ion circular trajectory is r such that:

$$zvB = \frac{mv^2}{r}$$

where B is the intensity of the magnetic field.

One way to scan the m/z range is to modify the intensity B of the magnetic field. This allows to use a constant radius r and to limit the size of the instrument. The mass analyzer selects the ions with mass-to-charge ratio equal to

$$\frac{m}{z} = \frac{B^2 r^2}{2V}$$

Magnetic sector mass analyzers are relatively slow because of the inertia of the magnetic field. In recent implementations, an array of detectors is used at the exit of the magnetic sector.

Advantages

- high reproducibility
- excellent quantitative performance
- high resolution
- high sensitivity

Limitations

- not compatible with MALDI ionization
- larger instrument size, higher cost
- slow scanning speed

3.3.3 Quadrupole mass analyzers

Principle Quadrupole mass analyzers consist of four precisely parallel metallic rods that are equally spaced around a central axis. Opposing rods are applied an oscillating radio-frequency voltage component and a direct-current component. One set is applied the positive part, and the

other set the negative voltage. There is no magnetic component in a quadrupole mass analyzer. Depending on the voltage frequency and direct current components, only ions with a selected m/z ratio have a stable trajectory and reach the exit of the mass analyzer.

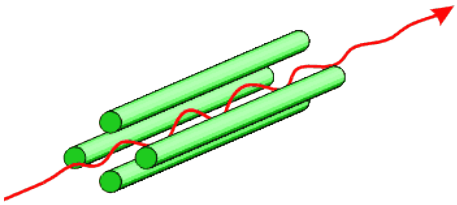

Figure 3.10: Quadrupole mass analyzer. Only ions with the selected m/z ratio traverse the space between the metallic rods.

Details The positively charged rods act as a high mass filter. Only the ions with a m/z ratio higher than a critical value are transmitted through the center of the quadrupole. Similarly, the negatively charged rods form a low mass filter. By combining the four rods, ions of a specific m/z ratio can be selected.

Quadrupole rods can have other functions besides their use as a mass filter. When using only the radio frequency component, the quadrupole acts as an ion guide. When using only the direct-current component, the quadrupole can be used as a focusing element is some designs (e.g. a linear ion trap).

A derivation of the working equations for a quadrupole mass analyzer is beyond the scope of this discussion, but it is based upon a second-order differential equation known as the Mathieu equation. A good description of the theory behind both quadrupole mass filters can be found in [MH89].

Advantages

- good reproducibility
- small instrument size, low cost
- efficient transfer of ions when used as an ion guide

Limitations

- limited resolution
- peak resolution is dependent on the mass of the molecule and must be tuned
- low compatibility with MALDI ionization

3.3.4 Time-Of-Flight mass analyzers (TOF)

Principle In a Time-Of-Flight mass analyzer, the mass spectrometer measures the time necessary for the ions to traverse a tube of predefined length and reach the detector.
More details can be found in [CLT01].

Details In a TOF mass analyzer, the peptide ions are accelerated to the same energy $E = zV$ where z is the charge of the ion and V is the potential difference used for acceleration. They acquire straight line motion at the speed

$$v = \sqrt{\frac{2E}{m}}$$

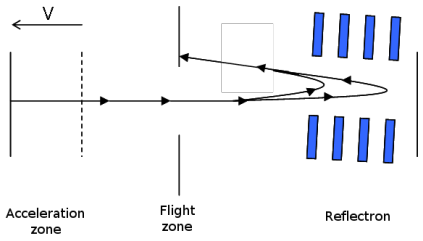

Figure 3.11: Time-Of-Flight analyzer with a reflectron.

where m is the molecular mass of the peptide.

Let L denote the length of the flight tube. Then m/z ratio of the peptide is a function of the time t needed to reach the detector.

$$\frac{m}{z} = \frac{2Vt^2}{L^2}$$

High ion velocities are reached in TOF instruments. On some instruments, special detectors with high sampling frequency are employed as described in Section 3.4.

Reflectron Higher resolution is achieved when the peptide ions are focused in the same location, in spite of their repulsive charges. This ensures that the same energy is provided to all ions. To further improve the resolution, TOF analyzers use mirrors that reflect the incident ions. Ions with higher energy penetrate further in the reflectron, fly a slightly longer path, and their time-of-flight is corrected.

Advantages

- fastest scanning speed
- well suited for pulsed ionization in MALDI
- high sensitivity
- highest practical mass range

Limitations

- requires pulsed ion acceleration
- high sampling frequency detectors have a limited dynamic range

3.3.5 Ion trap mass analyzers

Principle Ion trap mass analyzers focus the peptide ions in a confined space. Ions with a specific m/z ratio can then be selectively excited and detected.

A quadrupole ion trap is composed of a circular ring electrode and two end caps. Its principle of operation is similar to quadrupole mass analyzers, and the ion motion correspond to the same equations (see [Mar97] for more details). The peptide ions are first confined in the space between the electrodes, then ejected by adjusting the voltage applied to the end caps.

The ion trap has a limited capacity because of repulsive interactions between the peptide ions. Consequently, such mass analyzers have a limited dynamic range. To improve the focus, the movement of the ions can be dampened by introducing a neutral gas inside the trap .

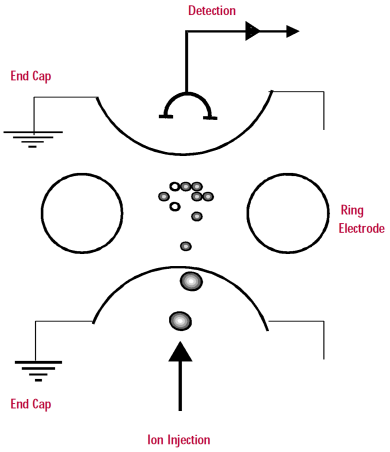

Figure 3.12: Ion trap mass analyzer. Ions are focused in the middle of the ring electrode.

To avoid saturation of the trap, a quick preliminary scan can be used to measure the ion beam intensity. The number of ions transmitted into the trap is then adjusted automatically.

Advantages

- High sensitivity
- robust, low cost instruments

Limitations

- limited dynamic range
- space charge effects inside the traps, and potential interactions between ions

3.3.6 Fourier Transform Ion Cyclotron Resonance mass analyzers (FT-ICR)

Principle The FT-ICR mass analyzer is an ion trap analyzer. However, instead of detecting ions that are ejected from the trap, the peptide ions are forced into a circular motion in the trap ("ion cyclotron"). As the ions turn around, they induce sinusoidal variations in the electrical potential of the detector electrodes. Fourier Transform is used to measure the rotational frequency of the ion cyclotron motion.
More details can be found in [MH02].

Details The FT-ICR ion trap is usually composed of six metallic plates forming a (cubic) cell inside a strong magnetic field. The three pairs of plates are used respectively for trapping the ions, exciting the ions and detecting the ion motion.

Trapping Ions are focused in the trap by the joint effect of the magnetic field and the trapping electrodes. As in the magnetic sector analyzer, ion motion in the magnetic field is circular in the plane orthogonal to the magnetic field direction. This constrains the ions in two dimensions. Trapping in the third dimension is obtained by applying electrical potentials to the trapping plates. As in the ion trap mass analyzer, focusing is improved by dampening the ion motions with a neutral gas inside the trap.

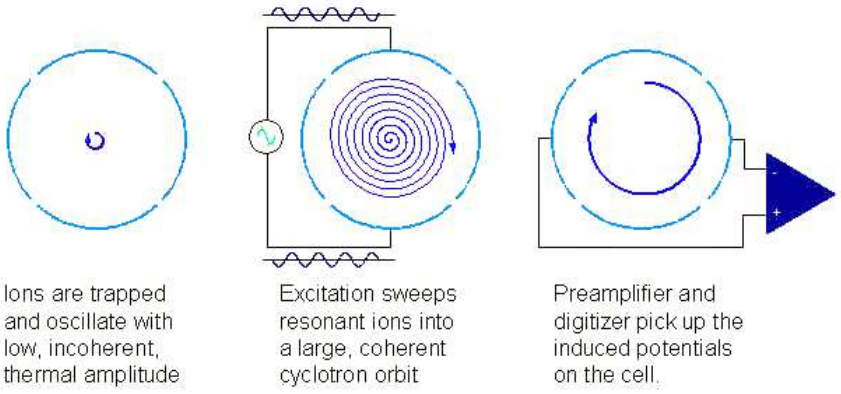

Figure 3.13: Three stages in FT-ICR mass analysis.

Excitation Without excitation, the radius of the cyclotron motion of ions inside the trap is too small to be detected. A rotating electrical field of strength E_0 and frequency ω_0 can excite the peptide ions with m/z ratio such that

$$\omega_0 = \frac{zB}{m}$$

where B is the intensity of the magnetic field. These ions are said to resonate with the electrical field.

The rotating field excites the ions which match is rotational frequency. When using a rotating field with multiple frequencies, the corresponding ions are all excited. A range of m/z ratios can be excited by using a "sweep" containing all the frequencies in the range. The excitation function is designed in frequency space, and Fourier transform is used to compute the actual form of the rotating electrical field. Consequently the FT-ICR is not used as a scanning device, although a scanning mode of operation is possible.

During excitation, ions with the same m/z ratio retain coherence, and move around in the trap in a tight packet. As the motion is coherent, the phase of the ions is the same and the motion is detectable. After excitation, the ion packet has a circular motion with radius:

$$r = \frac{E_0 T}{2B}$$

where T is the duration of excitation.

Detection The movement of the coherent packet of ions induces a detectable sinusoidal signal on the detector plates $Crzn\sin(\omega t + \phi)$ where n is the number of ions, ϕ is the phase and C is a constant. Fourier transform is applied to measure $\omega = \frac{zB}{m}$ for mass measurement.

Thanks to the superposition principle in electromagnetic theory, the sinusoidal signals induced by ion packets of different m/z ratios are additive on the detector electrodes. However, the capacity of the ion trap is limited, and, as with ion trap mass analyzers, preliminary scans are used to optimally fill the trap.

Influence of the magnet strength FT-ICR instruments are costly because of the superconducting magnets used to generate the constant magnetic field. A high intensity magnetic field benefits

- resolution, which is proportional to the number of rotations of the ion packet inside the cell
- better confinement and stability of the ions, so better trap capacity
- increased mass range

Advantages

- highest mass accuracy and resolution
- non destructive analysis, ions can be cooled then re-excited

Limitations

- limited dynamic range
- subject to space charge effects and ion interactions
- slow acquisition rate

3.4 Ion detection

This section deals with how the mass spectrometer measures the intensity of the ion beam. Except FT-ICR instruments, mass analyzers are scanning devices, i.e. ions corresponding to different mass-to-charge ratio are detected at different time points. We present here the different methods for ion detection used with scanning mass analyzers. Ion detection in FT-ICR instruments is described along with the FT-ICR analyzer.

3.4.1 Principle

The ion detector measures electrical currents corresponding to charge annihilation when peptide ions hit the detector plate. This is the principle of the simplest ion detector device called a Faraday cup. To improve the sensitivity of ion detectors the signal must be amplified.
More details on ion detection can be found in [Smi04, Wiz79].

3.4.2 Amplification

Mass spectrometers deal with relatively small numbers of peptide ions. For example, the capacity of an ion trap is on the order of thousands of ions. In mass spectrometry, the ion detector can detect single ion events thanks to amplification devices.

Amplification is based on secondary electrons produced by the impact of an ion on the detector. As indicated in Figure 3.14, the secondary electrons then produce additional electrons upon impact on the detector wall. As many as 100,000 secondary electrons can be produced by the impact of one ion.

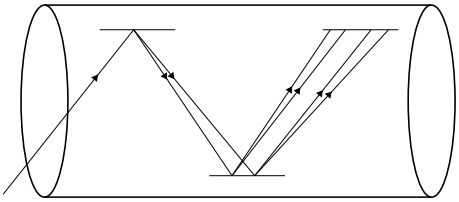

Figure 3.14: An electron multiplier.

In electron multiplication, the intensity of the measured signal is mainly dependent on the kinetic energy of the incident ion. This is why many of the acceleration methods in mass spectrometry provide ions with the same amounts of energy. In particular, triply charged ions do not produce three times more signal than singly charged ions, although the signal intensity is usually higher.

To increase the size of the detector, many electron multiplier devices can be coupled on a multi-channel plate as represented in Figure 3.15.

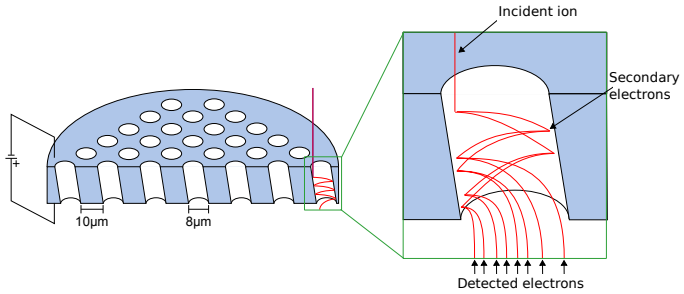

Figure 3.15: Multi-channel plates use an array of electron multipliers.

3.4.3 Saturation

An ion detector has a limited dynamic range, and there is saturation when the ion intensity is too high. The detectors commonly employed in LC/MS systems have a dynamic range around 10^4, that is to say it can measure intensities on a scale from 1 to 10,000.

To improve the dynamic range as well as the reproducibility of the measures, the mass spectrometer uses repeated measurements. A mass spectrum is the sum of a number of micro-scans, and that number can be adjusted in the parameters of the instrument.

3.4.4 Time-to-digital detectors

TOF instruments have a specific requirement as far as ion detectors are concerned. The typical flight time of an ion is around 90 ns in a 2m flight path. Consequently the ion detector mush have a high sampling rate in the time domain.

Time-to-digital converters trade dynamic range with a high frequency sampling rate. They provide a binary signal, either zero or one, that indicates whether an ion signal is present or not at a specific time point. These detectors cannot distinguish one ion event from several ion events. However, their sampling rate is very high, on the order of 100 GHz, so events separated by 10 picoseconds can be distinguished.

To obtain a mass spectrum, a large number of micro-scans are added in a TOF mass spectrometer, around 500. This does not compromise the scanning speed because each micro-scans only requires microseconds. The dynamic range of the detector is still limited to values between 0 and 500.

The dynamic range of the time-to-digital detectors is further limited by "dead time" effects. After a detected ion event, the detector is unable to react to other ion events for some time called dead time. Based on Poisson counting statistics, dead time correction procedures have been proposed (see for example [SZB94]):

$$\hat{\mathcal{I}} = -N \ln\left(1 - \frac{\mathcal{I}}{N}\right)$$

where $\hat{\mathcal{I}}$ is the corrected intensity, $\mathcal{I}$ is the measured intensity and N is the number of micro-scans.

Conclusion

Mass spectrometers can count individual peptide ions, and so are very sensitive instruments. That sensitivity is limited by the transmission efficiency of the ion source and of the mass analyzer, but also by the intrinsic dynamic range of the ion detector.

To count the number of peptide ions, the signal created by the incident ions is amplified with electron multiplication devices. This amplification factor can be integrated in the ionization efficiency of the peptide.

3.5 Instrumentation

A mass spectrometer is assembled from an ion source that provides charges to peptides and brings them into gas phase, a mass analyzer that separates ions based on their mass-to-charge ratio, and an ion detector that counts the number of ions produced.

These elements are usually interchangeable, and there are many combinations available for purchase. Mass spectrometers are designated with an acronym of the form ESI/FT-ICR/MS or nanoLC/MALDI-TOF/MS which indicates the type of each element. Ion sources, mass analyzers and ion detectors can be bought separately and assembled, but the manufacturers provide all-in-one solutions, driver software as well as maintenance contracts.

Each element has an impact on the acquired signals. The ionization efficiency is different according to the ion source, and peptides may or may not be detectable in some setups. The mass analyzer determines the resolution and accuracy of mass spectra, which is the main ingredient for identification of the proteins.

To conclude, we present a list of common mass spectrometers employed for LC/MS-based proteomics analyses, as well as LC/MS images acquired on these instruments. Most of these instruments actually combine two mass analyzers; the second one is used for peptide identification, as will be discussed in Chapter 4.

	Q-Q-TOF	TOF-TOF	Q-Q-Q	Ion trap	FT-ICR
Mass accuracy	good	good	medium	low	excellent
Resolving power	good	high	low	low	very high
Sensitivity		high	high	good	medium
Dynamic range	medium	medium	high	low	medium
Throughput	++	+++	++	+++	++
Electrospray	yes		yes	yes	yes
MALDI	optional	yes		optional	
Identification	++	++	+	++	+++
Quantification	+++	++	+++	+	++

Table 3.1: Performance of common mass spectrometers for LC/MS analyses. Adapted from [DA06].

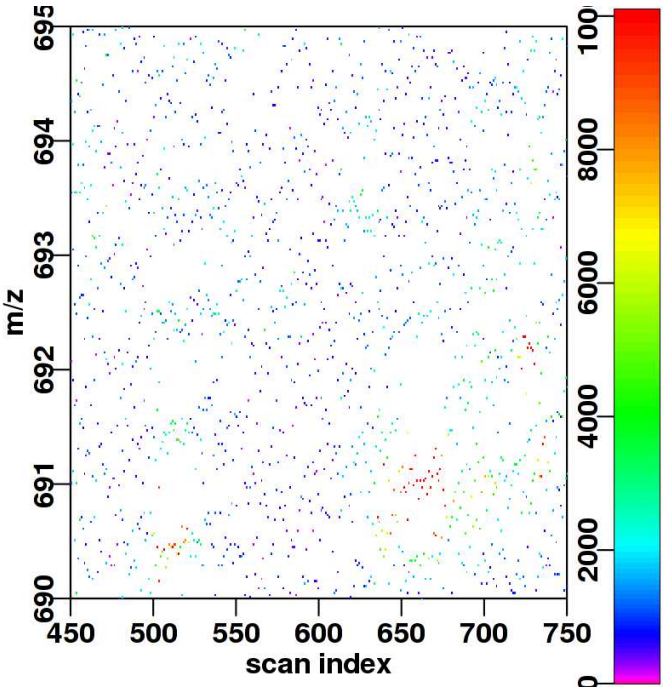

Figure 3.16: LC/MS image from a LTQ instrument (Linear Ion Trap mass analyzer). Figure 3.16, Figure 3.17 and Figure 3.18 were generated from the data set in [KEH+08] where a mix of 18 proteins are run on several instrumental platforms.

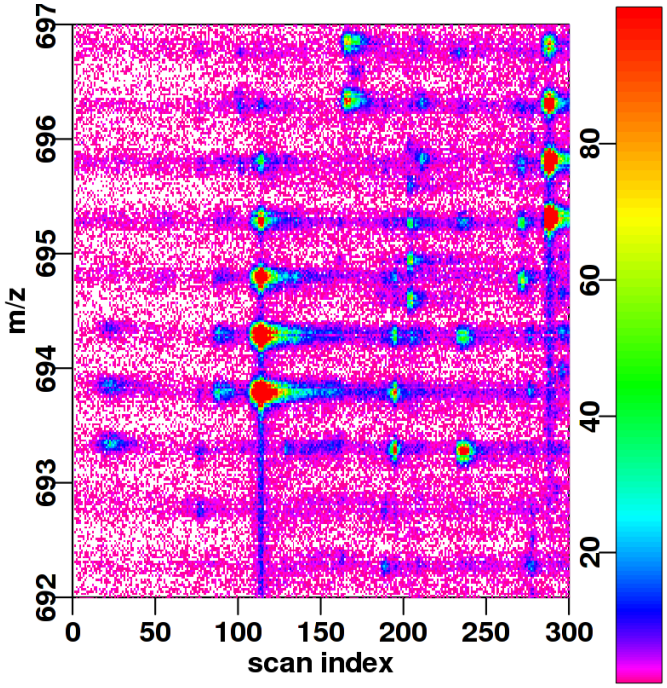

Figure 3.17: LC/MS image obtained from a Q-TOF instrument (TOF mass analyzer).

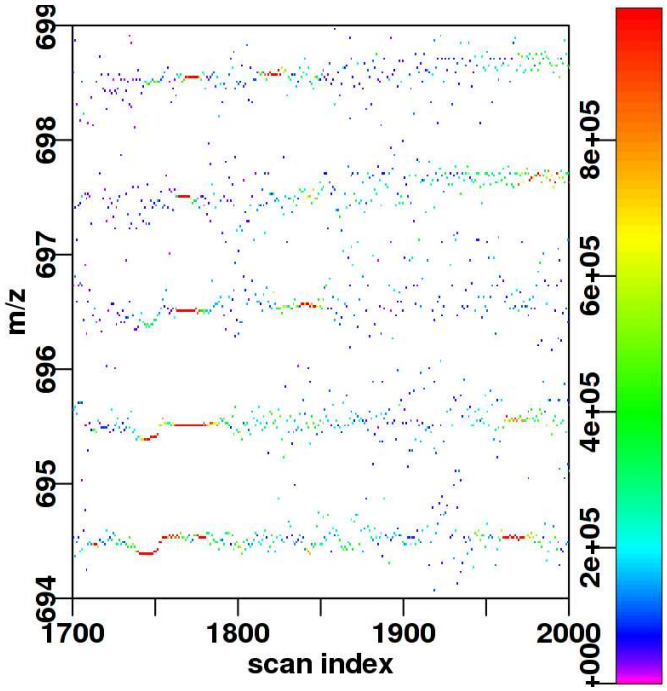

Figure 3.18: LC/MS image obtained on a LTQ-FT instrument (FT-ICR mass analyzer).

Chapter 4

Protein identification and quantification

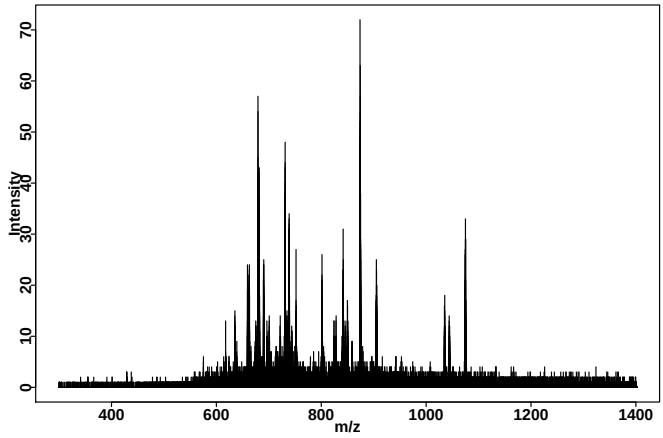

Figure 4.1: A mass spectrum.

Proteomics is an area of analytical chemistry which analyses protein samples. The primary objective is to identify all the proteins in the sample, and to measure their concentration. Mass spectrometry coupled with liquid chromatography has distinct advantages for proteomics applications:

- sensitivity, proteins with very low concentration can be detected and quantified.
- very low requirements on the minimum quantity of biological material. Single cells can be analyzed [DHM$^+$06].
- large-scale and high-throughput analyses can provide information (identification and quantification) on a large portion of the proteome.

A complex sample in proteomics contains more than a thousand proteins or peptides. For example, a lysate of yeast cells may contain up to 6,500 different protein species. These proteins are in a wide range of concentrations. Some estimates report a factor of 10^{10} between the concentration of abundant proteins and that of low abundance proteins [DA06]. Some cells may contain only a few copies of a given protein.

The current capabilities of LC/MS technology is about 1,000 proteins identified in a biological sample from a single experiment. When using extensive pre-fractionation procedures, this figure can be improved up to 10,000 proteins, at the cost of weeks of instrument time and sample manipulation. Quantitative measurements of the concentration up to four orders of magnitude have been reported.

In this chapter, we present the methods used to interpret LC/MS data, i.e. find the sequences of the proteins in the sample and their concentration. Protein identification methods are described in Sections 4.1 to 4.4, and protein quantification is described in Section 4.5.

These methods use technological developments that are specific to LC/MS and that take advantage of the biochemical properties of proteins. Basic notions in protein chemistry are presented in the Appendix, and we refer the reader to [Str88] for more in depth coverage.

4.1 Protein identification in MS spectra

Identification of proteins in mass spectrometry is based on the comparison between the observed mass-to-charge ratio of a molecule and the theoretical mass-to-charge ratio of the protein. The theoretical m/z ratio can be computed with very high precision based on the chemical formula of the molecule. In this section, we present identification of proteins based on isolated mass spectra. In particular, we show their limitations for large-scale analyses and motivate the use of methods based on molecular fragments.

4.1.1 Identification of proteins based on their m/z ratio

In a mass spectrum, the only available information for identification is the m/z ratio of a protein signal. Very high accuracy measurements of the m/z ratio have been used successfully for the identification of molecules of low molecular weight. Small molecules can be distinguished based on m/z alone because of the limited possible combinations of atoms. Structure determination is even possible because chemical bonds between atoms modify the molecular weight.

In proteomics analyses with tryptic digestion, the molecules considered have a molecular mass between 400 and 6,000 Da [1]. These peptides are much too heavy for identification based only on the molecular weight; this would require much more accuracy and resolving power than is available in current instrumentation. Moreover, the m/z ratio does not indicate the amino-acid sequence of the protein, as permutations in the sequence lead to the same molecular weight.

A careful study in [NMA+05] provides a numerical assessment of the required mass accuracy for m/z based protein identification. For example, in the $[400Da - 6,000Da]$ interval considered, there are 400,000 possible tryptic peptides in yeast cells, and 1.7 million in human cells. [NMA+05] estimate that the required accuracy is on the order of 0.1 ppm whereas the best instruments have an accuracy of 5 ppm.

Complementary information is necessary to supplement the m/z ratio of a protein. The Peptide Mass Fingerprinting method (PMF) presented in 4.1.2 uses the m/z values of peptides obtained by enzymatic digestion.

[1] TOF mass analyzers can measure m/z ratios up to 500,000 Da.

4.1.2 Peptide Mass Fingerprinting (PMF)

Peptide mass fingerprinting consists in identifying proteins based on the m/z ratio of the peptides obtained by enzymatic digestion. Trypsin is used to cut the protein into peptides, as in sample preparation (cf Chapter 2). The mass spectrum contains peaks at the m/z ratio of each peptide.

Example : Insulin B is a small protein with amino acid sequence

```
FVNQHLCGSHLVEALYLVCGERGFFYTPKT
```

After tryptic digestion, the following three peptides are obtained:

```
FVNQHLCGSHLVEALYLVCGER
FFYTPK
T
```

As indicated on Figure 4.2, the theoretical mass spectrum corresponding to Insulin B contains a peak for each of the fragments.

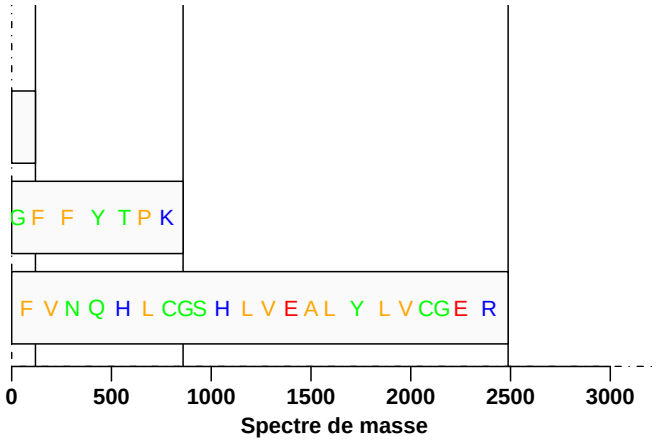

Figure 4.2: Theoretical mass spectrum obtained in PMF of Insulin B.

The observed mass spectrum is compared to the entries of a protein database. From the mass spectrum, a list of peaks and their m/z ratio is computed. For each protein sequence in the database, tryptic digestion is performed in silico, and the theoretical mass of the peptides is computed. A protein is identified if a sufficient number of peptides are found in the mass spectrum.

Limitations

- PMF is not applicable in conjunction with liquid chromatography because the tryptic peptides from the same protein are separated into different fractions and retention times. The method must be extended for application to LC/MS images.
- PMF is used to identify purified protein samples with no more than two or three concurrent proteins. Otherwise, there are too many peaks in the mass spectrum, many possible interpretations and a high possibility of a false positive identification.
- Identification of proteins is dependent on the database. Proteins that are absent from the database cannot be identified, but exhaustive databases can be constructed based on the genome sequence (see Appendix).

Peptide Mass Fingerprinting was developed for analysis of proteins separated on a gel. The proteins in each spot are identified by mass spectrometry. Gel-based proteomics is still in use because the gels have a higher separation power than liquid chromatography. However, gels are more difficult to interface with mass spectrometry. For instance, proteins need to be excised from the gel; the method has low throughput because each gel spot is analyzed separately.

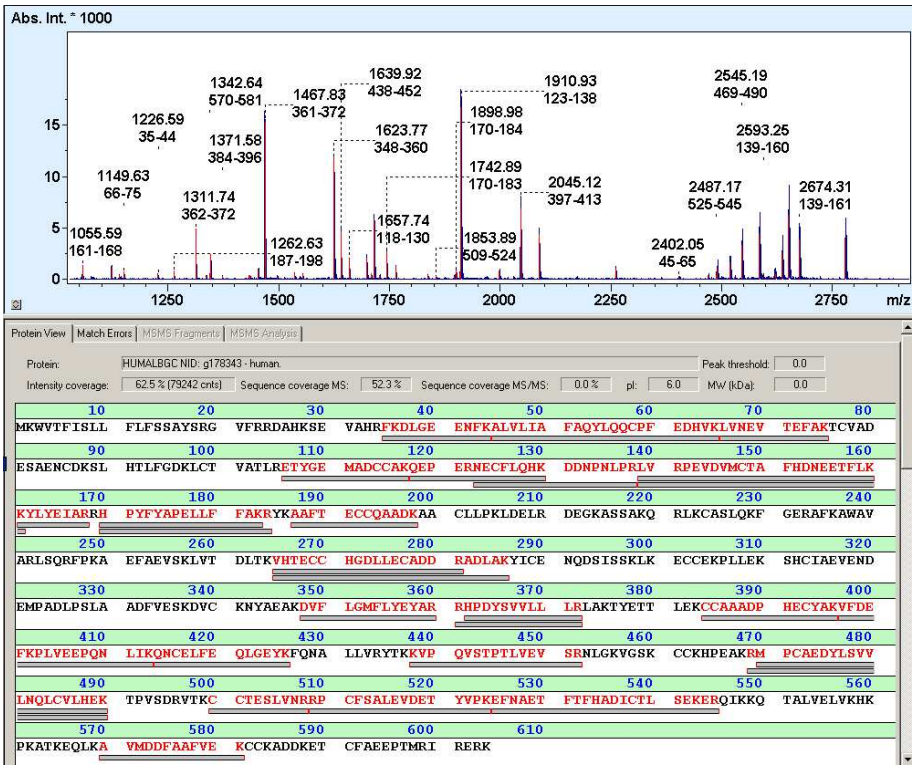

Figure 4.3: Identification of human serum albumin with PMF. The top panel displays the mass spectrum acquired on a MALDI-TOF platform. The bottom panel represents the amino acid sequence of albumin. The letters highlighted in red and the gray bars indicate the peptides that are found in the mass spectrum.

Summary

Identification of molecules based on their molecular weight is only possible for small molecules and with very high accuracy instruments. This method is not applicable for proteomics analyses.

Proteins need to be identified based on several m/z measures. Peptide Mass Fingerprinting is an identification method based on the molecular weight of tryptic peptide fragments of the protein. It is widely used in proteomics in conjunction with gel-based separation of the proteins.

LC/MS-based protein identification methods use supplemental information. Tandem MS methods require an additional fragmentation experiment inside the mass spectrometer and are described in Section 4.2. The methods presented in Sections 4.3 and 4.4 use the retention time of proteins provided by the liquid chromatography separation.

4.2 Protein identification in MS/MS spectra

Tandem MS or *MS/MS* identification is based on peptide ion fragments generated inside the mass spectrometer. Most mass spectrometers used for LC/MS analyses have a tandem MS mode of operation. This identification method requires adequate instrumentation, but provides the most reliable results at the moment. A very detailed discussion of software and methods for MS/MS interpretation can be found in [HMA06, JDTP05].

In MS/MS mode, two MS spectra are successively acquired. The first mass spectrum (the MS spectrum) is a standard analysis the contents of the sample as a function of the m/z ratio. It is used to select m/z values of interest. Once a m/z value is chosen, the corresponding molecules are fragmented in the mass spectrometer, and the second mass spectrum (the MS/MS spectrum) is acquired. If necessary, some instruments (in particular trapping instruments) can perform several stages of fragmentation and MS analysis.

A mass spectrometer cannot usually acquire MS spectra and MS/MS spectra in parallel. The instrument alternates between the two modes, so MS/MS identification requires instrument time. While performing MS/MS, proteins continue to elute in the liquid chromatography column, and the obtained droplets are diverted from the mass spectrometer. Consequently, performing MS/MS fragmentation reduces the available data for quantification (lower sampling rate of the LC/MS image) and is limited in the number of identifiable peptides (co-eluting peptides are lost).

Identification is composed of the following steps:

- selection of the parent ion for identification
- fragmentation of the parent ion
- acquisition of the MS/MS spectrum
- interpretation of the MS/MS spectrum.

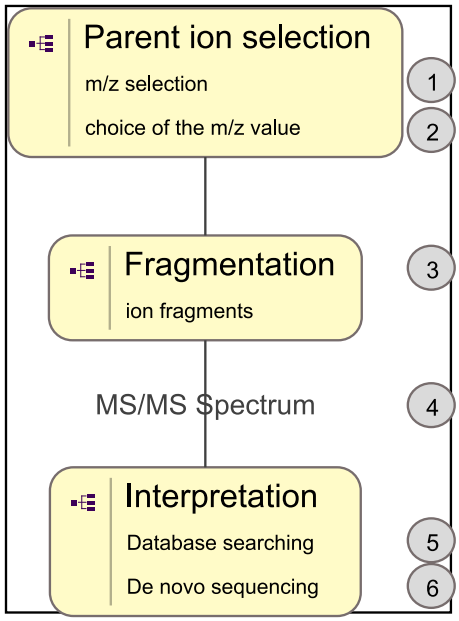

Figure 4.4: Analysis pipeline in MS/MS based peptide identification. Each step is discussed in the sections indicated with the circled numbers.

4.2.1 Selection of the parent ion

Parent ion selection consists in extracting a homogeneous set of ions corresponding to a single peptide species. This corresponds to selecting ions at specific retention time and m/z ratio in the LC/MS image.

Selecting a specific retention time is easy. In electrospray ionization, only one droplet is available for MS analysis at any given time, so parent ion selection is specific to that droplet. In MALDI ionization, the retention time is selected by focusing the laser beam on a specific spot of the MALDI plate.

Selection of the ions with a specific m/z ratio is performed by the mass analyzer. Scanning mass analyzers are easy to use as m/z ratio filters. It suffices to set the filter to the desired m/z ratio, and all other ions are removed. In ion trap mass analyzers and FT-ICR mass analyzers, unwanted ions are selectively ejected from the trap.

The time-of-flight analyzer is more tricky to operate as a mass filter because all ions follow the same path, and m/z separation is based on their time of arrival. A special type of electromagnetic gate ("Bradbury-Nielson gate" or "Time Ion Selector" [ZBR+95]) is open briefly and only the ions of interest are allowed through. The Time Ion Selector must be operated at very high speed.

In trapping mass analyzers, peptide ions with the selected m/z ratio are kept inside the trap. In scanning mass analyzers, only the selected peptide ions reach the exit of the mass analyzers.

4.2.2 Criteria for choosing a parent ion

During large-scale analyses, the objective is to identify as many proteins as possible and all proteins are potentially interesting. Consequently, there is no a priori or preferred m/z value to be selected for MS/MS identification. The natural strategy is to consider the most abundant m/z values for fragmentation.

Parent ion selection is implemented in the driver software provided by the instrument manufacturer. Currently, only the natural strategy is available for selecting the parent ions. It implements the following rule:

> MS/MS spectra are acquired for each MS spectrum, and the n most intense ions are selected, with n a user-defined parameter, usually from three to five.

In LC/MS analyses, most peptides are present in several adjacent fractions. When one peptide has a high intensity in a given mass spectrum, it is likely to have a high intensity in the following MS spectra as well. To refrain from selecting the same peptide over and over, mass spectrometer software implement exclusion lists. These correspond to the following rule:

> After selection for fragmentation, a m/z ratio enters the exclusion list for t seconds. During that time, the m/z ratio cannot be selected for fragmentation. t is usually on the order of one minute.

There are also inclusion lists which work in a similar way.

The rationale for choosing abundant ions is that MS/MS spectra require a sufficient signal intensity for interpretation. The drawback is that high intensity ions correspond to proteins that are usually well-known, whereas potential biomarkers are excepted in low-intensity signals.

Only a limited number of ions can be selected for MS/MS fragmentation. In electrospray ionization, there is a tradeoff between the number of MS/MS spectra and the number of MS spectra. During the acquisition of MS/MS spectra, other eluting peptides are lost. This is not a problem in MALDI ionization because MS/MS spectra can be acquired off-line. However, the separation power in MALDI is typically lower because less fractions are used.

4.2.3 Fragmentation

There are many fragmentation methods. We will only present Collision Induced Dissociation (CID) which is the most widespread. In depth coverage of fragmentation methods can be found in [SV04].

Collision induced dissociation fragments peptide ions by first accelerating them, then colliding the ions with a neutral gas. Dissociation of the amino acid sequence occurs between the amino acids. The parameters are set such that from each peptide ion, there is only one fragmentation event. Consequently, only prefixes and suffixes of the protein sequence are formed.

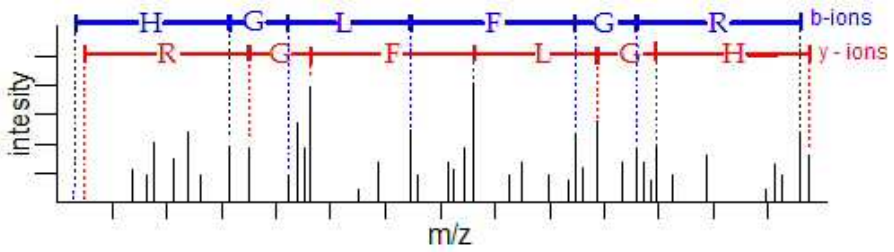

Figure 4.5: De novo interpretation of an MS/MS spectrum of HGLFGR. Among the observed peaks, b-ions correspond to prefixes of the peptide sequence, and y-ions correspond to its suffixes.

Example Consider the tryptic peptide `HGLFGR`. After fragmentation, we expect the following fragments as in Figure 4.5:

```
H    GLFGR
HG    LFGR
HGL    FGR
HGLF    GR
HGLFG    R
```

The charge of the parent ion is transmitted to its fragments at random. When the parent ion is singly charged, only one of the fragments receives the charge, and the neutral fragment is not detected. When the parent ion holds more than one charge, all charge configurations are possible for its fragments.

In scanning mass analyzers, ions are accelerated for mass analysis, and it suffices to introduce the collision gas at the exit of the mass analyzer. In trapping mass analyzers, the ions are excited and undergo a high velocity cyclotron motion. The collision gas is introduced in the trap, and the fragments are collected inside the trap.

Alternative fragmentation methods like Electron Transfer Dissociation (ETD) are significantly different from CID because not exactly the same fragments are produced. ETD has a lot of potential to replace CID because it is better suited to large molecules and because it produces a greater variety of fragment ions, which benefits the interpretation of the MS/MS spectrum.

4.2.4 Acquisition of MS/MS spectra

MS/MS spectra are obtained much like MS spectra. In scanning mass analyzers, the analyzer is used for parent ion selection and only the selected ions exit the mass analyzer. Parent ions are fragmented at the exit, and a second mass analyzer is used to acquire the MS/MS spectrum. This is why two mass analyzers are implemented in current instrumentation as shown in Table 3.1 (see Section 3.5, page 46).

With trapping mass analyzers, only the selected parent ions are kept in the trap. After fragmentation, the fragment ions remain in the trap, and the same mass analyzer can be re-used for acquisition of MS/MS spectra.

Special cases

- Triple quadrupole instruments (Q-Q-Q) have three mass analyzers, but only two are used for m/z separation. The second quadrupole is used as a collision chamber and operated as a linear ion trap or ion guide. MS/MS acquisition is performed by the third quadrupole.
- FT-ICR mass analyzers are relatively slow and are not usually used for MS/MS in LC/MS analyses. Instruments like the LTQ-FT, actually combine two mass spectrometers: a linear ion trap (LTQ) and the FT-ICR instrument. In MS/MS operation mode, everything is carried out by the LTQ, from parent ion selection to MS/MS acquisition. The FT-ICR mass analyzers acquires a high resolution MS spectrum in parallel.

4.2.5 Interpretation of MS/MS spectra based on a protein database

MS/MS spectra are usually interpreted by comparison with a protein database. Although the fragments are different, the conceptual ideas are very similar to Peptide Mass Fingerprinting. For every entry in the protein database, the interpretation software predicts the ions fragments and their m/z ratio based on the chemical formula. Theoretical spectra are compared to the observed mass spectrum and the best match is reported.

Limitations

- As in PMF identification, only the proteins in the database can be identified, but exhaustive protein databases can be constructed from the sequencing of the genome.
- Only a limited number of MS/MS spectra are of good enough quality to be interpreted. This number varies between 20% to 70%.
- Only a limited number of parent ions are selected for MS/MS identification, and many low-intensity signals are not identified.
- MS/MS identification requires a lot of computational time because of the database search.

Assignment of the correct identity depends strongly on the scoring function used to match the experimental and theoretical spectra. In these comparisons, the intensity of the peaks in the MS/MS spectrum is not used, as it cannot be predicted from the protein sequence. The first scoring methods relied on cross-correlation (Sequest [EMY94], Mascot [PPCC99]). Newer methods cast the score into a probabilistic framework, and report an estimate of the p-value (i.e. the probability that a given score occurs by chance), for example ProteinProphet [NKKA03]. A review of these methods can be found in [SCYI04].

Only the peptides in the database that match the molecular mass of the parent ion are considered. As a consequence, a higher mass accuracy speeds up database searches, and improves the confidence in the identification results.

To further speed up database searches, MS/MS spectra are preprocessed, and low-quality spectra can be left out. Other strategies define sequence tags which are sets of peaks that can be attributed to a small sequence of amino acids. Only the peptides in the database that contain the sequence tag are compared to the MS/MS spectrum.

4.2.6 De novo interpretation of MS/MS spectra

De novo interpretation of MS/MS spectra infers the sequence of the parent peptide without a protein database. This requires higher quality MS/MS spectra because the search space is much larger; it contains all possible combinations of the 20 amino acids.

One approach for de novo interpretation is to create a pseudo protein database containing all possible sequences with matching parent ion mass. Each possible sequence is then scored against the observed MS/MS spectrum. This approach suffers from combinatorial complexity and is seldom used.

Current approaches look for matching pairs of peaks in the MS/MS spectrum that differ by the mass of one amino acid. These pairs are attributed to pairs of fragments, one small and one with the added amino acid. The amino acid sequence of the peptide is reconstructed step by step.

When the sequence information is incomplete, identification can be achieved by searching a protein database with the sequence elements found by de novo interpretation.

Limitations

- Even when all the fragment ions are observed, some parts of the peptide sequence can remain unknown. This is because amino acids like leucin and isoleucin have the same molecular mass. Some pairs of amino acids also have matching mass, see Table 2 in [JDTP05] for a list of such problems.
- De novo interpretation requires more computational time and higher quality MS/MS spectra than database interpretation. De novo is usually less powerful.

4.3 Identification of proteins in LC/MS images (introduction)

4.3.1 Context

MS/MS identification has two main disadvantages. It can only identify a small subset of the signals in a LC/MS image because of the limited number of MS/MS spectra, and retention time information provided by the LC separation is ignored.

Protein identification based on aligned LC/MS images trades the instrument time devoted to MS/MS analyses for higher sensitivity in MS spectra and higher sampling rate. The methods reviewed in Section 4.4 are applied after correction of retention time differences between different data sets.

The contents of Section 4.4 have been extracted from my work in [VLTTK+08]. To introduce the subject, we present in detail the Accurate Mass Tag approach (AMT) described in [ZMQS06] which is representative of other approaches in the field.

4.3.2 The Accurate Mass Tag approach for identification

The AMT approach combines the high mass accuracy of an FT-ICR mass spectrometer with a reproducible nano-LC separation. In that case, protein signals can be identified simply based on their m/z ratio and their retention time.

Identification based on the position of the signal in the (retention time, m/z ratio) plane puts strong requirements on the LC/MS platform. According to [NMA+05], even if the retention times vary by no more than 1% and the m/z accuracy is better than 1 ppm, half of the tryptic peptides in the human proteome cannot be uniquely identified.

When the complexity of the proteome is lower, as in the case of yeast samples for example, the requirements for identification are also relaxed. The AMT approach implicitly uses the fact that not all of the potential peptides can effectively be detected by the mass spectrometer. This leads to the reduction of the practical proteome in consideration, and has been successful for microbial and mammalian systems.

The AMT approach consists in two stages:

- First, a series of controlled experiments are used to build a database of Potential Mass and Time (PMT) tags. MS/MS identification is used to determine the sequence of the peptide signals in each experiment, and for each signal, its m/z ratio and retention time are corrected with m/z calibration and retention time alignment as described in Chapter 6. The PMT tag database contains a list of peptides, with their typical m/z ratio and retention time and their standard deviation.
- Second, the biological sample of interest is analyzed without MS/MS identification. The peptide signals detected in the LC/MS image are matched to the PMT tag database after m/z calibration and retention time alignment.

As in the AMT approach, the methods reviewed in Section 4.4 build on the assumption that peptide signals in LC/MS images can be matched based on their position in the image. To enable this, it is assumed that retention times have been corrected.

4.4 Peptide Identification in aligned LC-MS images

In this section, we discuss peptide identification in aligned LC-MS images, a major application area for LC-MS image alignment. This approach is complementary to the classical, direct, protein identification by LC-MS, where peptides and proteins are identified from fragmentation spectra generated in the same experiment.

In the classical approach, spectra are identified by searching each spectrum against a protein or DNA database, by de novo interpretation of the spectrum without a database, or by comparison with previously recorded MS/MS spectra [YWH+05]. When applied to complex samples, a significant limitation of this approach is that only a small fraction of all peptides can be selected for fragmentation by the mass spectrometer, as the number of possible MS/MS scans in a single experiment is limited. Typically, only the most intense MS peaks are selected for fragmentation, leaving less abundant peptides and proteins unidentified [LYK+05]. This limitation can be countered by multiple LC-MS analyses on the same sample or additional fractionation steps [EF07], but only at the cost of a major decrease in throughput.

Advances in elution time alignment, as surveyed in Section 6.4, have enabled the alternative approach, discussed here, called "propagation of peptide identity" [JML+06]. In aligned LC-MS images, a feature with an associated identification may have similar coordinates as another,

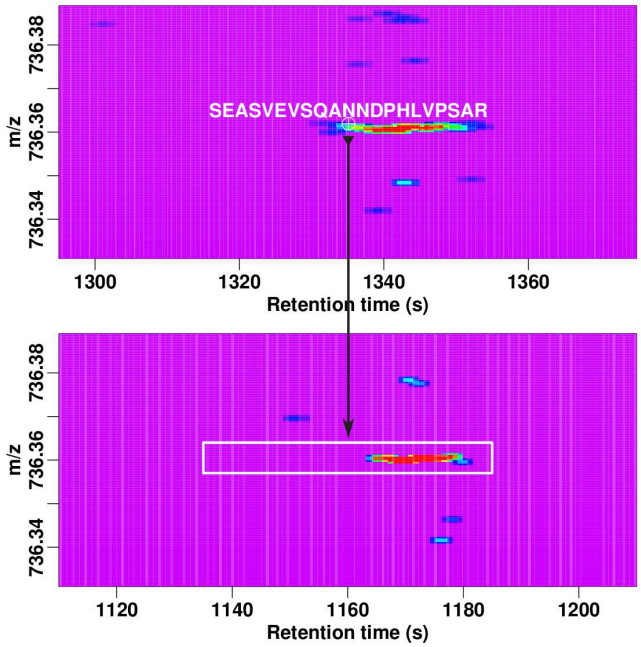

Figure 4.6: Example of peptide identity propagation. The figures show a small window of the data presented in Fig. 6.8. After alignment of the two images, the MS/MS identification in the upper image (cross-hair) can be propagated to corresponding MS features in the other image. A white rectangle denotes the valid range of putative corresponding MS features.

unidentified features in another image. This proximity in m/z and elution time coordinates may be interpreted as evidence that the previously unidentified feature was generated by the same peptide underlying the previously identified feature. As the availability of highly accurate mass spectrometers and data readily increases, the opportunities to apply this principle multiply, and many methods have recently been published that exploit it.

In a theoretical study, Norbeck et al. [NMA+05] explore the power of peptide identity propagation in different experimental scenarios. They find that complex protein mixtures give rise to peptides with very similar m/z values and (predicted) elution times. For robust identification, a combination of high m/z accuracy and highly accurate normalized peptide elution time (accurate to below 1%) appears to be essential for peptide identity propagation, particularly in complex mammalian mixtures.

While we discussed the technical aspects of LC-MS image alignment in Section 6.4, we survey in this Section their applications to peptide identity propagation. Section 4.4.1 covers those approaches for image-based peptide identity propagation, which operate directly on chromatographic information after image-based alignment, as discussed in Section 6.4.1.1. Section 4.4.2 covers feature-based approaches – those approaches in which LC-MS image features have already been detected and aligned, using the methods discussed in Section 6.4.1.2. Section 4.4.3 discusses validation approaches, a particularly important aspect for a new experimental large-scale approach.

4.4.1 Image-Based Peptide Identity Propagation

Image-based peptide identity propagation is based on the proximity of pixel features across aligned images. Radulovic et al. [RJR+04] detect rich features by grouping adjacent neighboring

image pixels with high intensity into a single peak, described by a specific m/z and elution time range (which they call pamphlet). Larger pamphlets (with more pixels) are processed before smaller ones. If new pamphlets overlap with more than one defined group, this group is split and reorganized, such that only a single group is retained from the successive addition of peaks.

Each high-confidence peptide identification, derived by searching MS/MS spectra against protein-sequence databases with SEQUEST [EMY94] or STATQUEST [KRR+03], is mapped to a data pamphlet. As a measure of the reliability of peak detection, the mapping is tested for sufficient peak-to-peptide overlaps. The alignment approach is based on a measure of similarity between two data sets to calculate the percentage feature (pixel) overlap. It exhaustively searches for a global optimum using a large set of pamphlets.

Wang et al. [WTF+07] model the contribution of one unit of abundance of a single peptide to an LC-MS image. This contribution is characterized by the monoisotopic mass and elution time range, and the collection of all possible such contributions is represented in a peptide element library. This library is used to obtain an alignment by matching the peak features in each profile to this common library: peak features from different profiles matched to the same peptide element are features representing the same peptide and should be aligned.

Given a library, a robust regression scheme is used to find the peptides that are present in each scan, together with their quantity. As a full peptide element library is usually not available, a subset is constructed that minimizes a loss function. This subset consists of a set of peptides deemed present in all input experiments.

The performance of the alignment is assessed both in terms of efficiency of recognizing features corresponding to the same peptide across samples, and the rate of false feature alignments (incorrectly aligned features corresponding to different peptides). The correlation between the intensities of aligned features between two replicate profiles is used as a measure of alignment quality: the more falsely aligned pairs, the less correlated the intensities of aligned features tend to be.

4.4.2 Feature-Based Peptide Identity Propagation

The methods described in this Section are based on preprocessed data, in which image data has been converted into discrete features. In the AMT-approach by Smith et al. [SAL+02], an early representative of identity propagation, preprocessed LC-MS/MS data are stored into a database of trusted features which further serves as the basis for the identification of the features contained in newly acquired LC-MS images.

The first stage of the AMT approach is based on a number of LC-MS/MS experiments that focuses on obtaining high-confidence peptide identifications with associated elution times. A normalization method is used to cluster and normalize run-to-run variations in absolute peptide elution times. The peptide identifications together with the normalized elution times are stored as MT tags in an accurate mass and time tag (AMT) database. The second stage of the AMT approach focuses on exploiting this database for identifying peptides from high-resolution LC-MS experiments.

A set of software tools (called VIPER [MTJ+07]) has been implemented to discover, align and match LC-MS image features to the AMT database. The features in a newly obtained LC-MS image are sorted by mass and broken down into mass bins. Each identified feature in an LCMS image is assigned a median mass, a NET and an abundance estimate. The elution times of all LC-MS image features are calibrated using either a linear alignment function or the LCMSWARP algorithm [JMP+06]. Features are matched to the closest AMT tag, by matching the monoisotopic mass and calibrated elution times of the features to those peptides in the database, and assigned its sequence. Ambiguous identifications, the normalized mass and elution time values of which are beyond the threshold value determined by the scoring scheme, are discarded in the process.

While this approach allows very efficient peptide identification without the need for routine MS/MS, highly reproducible separations and accurate MS measurements are prerequisites for

this method. Smith et al. [SAL+02] employ ultrahigh pressure capillary LC combined with FT-ICR MS. A notable restriction of the method is that only peptides with previous identification in the training set can be identified in subsequent experiments.

Other approaches more generally combine trusted MS/ MS information in multiple LC-MS experiments, which enables the exploitation of newly obtained CID-information that can serve to identify peptides in new, as well as in previous experiments. A common basis for most methods is a group of LC-MS image features that approximately match in m/z and elution time. If one or more peptides have previously been identified (by MS/MS identification, a match to the AMT database, or predicted elution time), their identity may be propagated to the remaining peptides. Statistical models can be used to assess the significance of the propagation step: the likelihood to observe a significant difference in m/z or elution time value [JMP+06], or similarly, a feature overlap- score [RJR+04, MRS+07] or an m/z p-value [JML+06] can be computed.

The PEPPeR system by Jaffe et al. [JML+06] is another sophisticated platform for feature-based analysis. In a first stage, features in MS spectra are calibrated by mapping confidently identified peptides to features identified by the Mapquant software [LSJ+06], using relaxed m/z ($\pm$ 25 ppm) and elution time (0.3 min) tolerances. A least-squares quadratic fit recalibration of m/z values is based on the identified features and leads to more stringent tolerances for the residual m/z errors (typically, $\pm$ 5 ppm). These tolerances are then used to remap identified peptides to features.

The features with an assigned identity (termed landmarks) serve as the basis for the second stage, in which new landmarks are sequentially identified by propagation of identities across experiments, before any alignment takes place. To identify a feature in a given spectrum by identity propagation, a reference mass spectrum is selected in which a landmark peak within the above stringent tolerance is present. The heuristic rules that allow selecting a best possible reference spectrum include maximization of overlap of previously established landmarks between the two spectra. To decide whether the peptide identity is being propagated, a score is computed and compared against a global threshold, which is chosen against a null model to control the number of randomly occurring identity propagations.

In the third stage, a coarse elution time alignment is computed by a quadratic fit based on landmark peptides across all experiments. To deal with experimental outliers, robust statistics are employed.

In the fourth stage, the m/z axis is partitioned into strips such that each strip represents the spectrum features belonging to one or more peptides. This speeds up the remaining computations, and makes the final stage parallelizable. Independently for each strip and each charge state, the corresponding features are clustered, using a Gaussian mixture model, which assumes that m/z and elution time vary around their fixed means. The number of clusters (hypothesized peaks) is determined by the Bayesian information criterion. After post-hoc merging of clusters that are too close by in terms of the previously determined m/z and elution time tolerances, a final peptide-identity propagation step is performed among the features within each final cluster.

After a feature-recognition step, Jaitly et al. [JMP+06] divide an LC-MS experiment into N segments, and a reference experiment into 3N segments. For alignment, a similarity score is computed between each segment of any LC-MS image, and each segment of the reference experiment. The scoring method is based on the variability of mass and elution time that quantifies the similarity between parts of the LC-MS images and LC-MS/MS datasets. The similarity score is computed by dynamic programming to minimize the distance between possible common features (weighted using a Gaussian model).

Individual features are paired explicitly afterwards, by first transforming the elution time of each feature into a normalized elution time corresponding to the reference experiment. Next, for each feature f, the closest feature g in the reference set is found using the lowest Mahalanobis distance.

The match score between the set of feature matches f and g corresponds to the log-likelihood of observing the mass and NET differences given a corresponding bivariate normal distribution (for mass and elution time errors, which are assumed to be independent, for each individual feature). Variability in measured elution time is modeled by a Gaussian distribution with small variance, and allows nonlinear warping on a larger scale.

A tool for high-resolution LC-MS based protein profiling – SuperHirn – has been developed by Mueller et al. [MRS+07]. MS/MS peptide identifications from a database search (using SEQUEST [EMY94], and PeptideProphet [KNKA02]) are associated with detected LC-MS image features, based on correspondence between elution time, charge state, and the theoretical mass of the peptide sequence. In cases where more than one peptide can be assigned to a feature, the peptide identification with smallest m/z and elution time difference is selected.

An iterative strategy (sketched in Section 6.4.1.2) uses identified features for global alignment. Common features of two LC-MS images are identified and associated within wide m/z and elution time tolerances (0.05 Da/ 5 min). These associations are used to compute a feature overlap score (which depends on the number of common features in two LC-MS images), to assess the overall similarity of two LC-MS images. The similarity analysis is used to construct a tree that determines the order in which the pairwise alignment between LC-MS images is successively extended to a multiple LC-MS alignment. Aligned features from all LC-MS images are stored in a repository map called "MasterMap", together with their corresponding MS/MS identifications and intensity profiles. This map includes all detected precursor ions, both known and unknown, and their normalized intensities across the range of biological samples.

SuperHirn searches for common features in another image within defined m/z and elution-time tolerance windows (0.01 Da and 0.5 min). MS/MS identifications are propagated from one aligned feature to another if the latter has not been already identified by a high-quality peptide assignment. Subsequently, protein profiles are computed by

1. the construction of LC-MS image feature profiles,
2. the detection of naturally occurring profile trends by K-means clustering and
3. the evaluation of constructed peptide and protein profiles by comparing them to theoretical abundance profiles.

Specifically, features with identical molecular mass, but different charge states, are grouped to form a peptide profile that consists of one quantity for each LC-MS experiment. MS/MS identifications are used to propagate identity to previously unidentified features. Peptides that belong to the same protein are expected to lead to correlated profiles. To evaluate the correlation of a peptide profile to a target profile, profile scores are calculated to build a distribution of scores. This distribution is modeled by a bivariate Gaussian mixture model, which is then used to compute a probability of a true correlation.

Rinner et al. [RMH+07] report an interesting use of SuperHirn for the identification of proteins in complexes containing the transcription factor FoxO3A in human cells. Experimental data are generated from co-immunoprecipitation dilution series experiments, which are designed to induce a defined peptide profile specifically for the peptides belonging to the proteins of interest. By identifying specifically those peptide profiles that follow the expected peptide profile, LC-MS image features can be selected for targeted MS/MS, leading to the high-confidence identification of FoxO3A complex members.

4.4.3 Evaluation of Protein/Peptide Identification

Due to the irreproducibility of the LC column, identification of peptides by searching in databases or by comparing with previously identified features does almost never produce exact matches: peptide identifications are not certain. Commonly, database search algorithms are used for single experiments to calculate the probability of obtaining random fragmentation (e.g. Mascot

[PPCC99]) or use heuristics (e.g. SEQUEST [EMY94]), which are then converted into probabilities for accurate peptide identification [KNKA02].

The sensitivity and specificity of peptide sequence identifications from MS/MS data can be significantly improved by performing the same search procedures on data recorded from multiple technical replicates [LYK+05]. Replicates are obtained by using different instrumentation and/or technical procedures, and more than one search algorithm can be used [CM05, EHFG05]. However, as thousands of measurements are taken from multiple experiments, the major challenge concerns to find an effective compromise between sensitivity of detection and rates of false positive error.

Statistical methods have been developed for evaluation of identification results, such as PeptideProphet [KNKA02] and STATQUEST [KRR+03]. These algorithms use score functions to determine the probability of each putative peptide match by making use of a collection of variables that are obtained from search engines, such as SEQUEST [EMY94]. By fitting the parameters to a certain distribution function, correct and incorrect matches can be distinguished.

With rapid advances in large-scale proteomics technology, the need became apparent for methods that improve sensitivity and control of false protein identifications by integrating the results of multiple experiments. These repeat experiments allow the detection of consistently expressed peptide features in different experiments. The use of a series of different samples containing protein mixes with known concentration ratio, and optionally dilution series, are commonly used [JML+06, WZL+03]. It is to note that database search methods frequently use different assumptions and can result in markedly different sets of protein identifications [SOB+06]. The major problem is that many protein identifications have low reproducibility, and only abundant proteins are reliably identified [CM05].

Peptide MS/MS spectra interpretation can be problematic when low-intensity non-peptide species are selected for fragmentation, the peptides being examined are not present in the queried sequence database, or the MS/MS spectra are not of sufficient quality for definitive interpretation. For these reasons, some degree of ambiguity is usually associated with each peptide identification, and previously required either manual inspection or the application of scoring filters based on how the filters performed on an often smaller and unrelated training dataset. The target-decoy search strategy provides a way to alleviate manual inspection and improve the search results both in throughput and accuracy, and is the best available option for analyzing largescale studies [EG07]. This decoy database is usually created by reversing [PET+03, Han03], or randomizing [PET+03, WAM+05] the protein sequences of the real protein database.

In the analysis of related samples, a significant difference between the different samples is reflected by an increase in the CV, characteristic for an increase in the SD to a normalized m/z or elution time parameter of the detected features. This, in turn, reflects significance from the point of view of biological species-content. This stresses the necessity for analytical methods that can integrate data from multiple experiments, which will likely increase, as common standard datasets emerge that facilitate the exchange of proteomic datasets from diverse sources [KEH+08].

4.5 Quantification

In mass spectrometry, quantitative measurements are deduced from the intensity of the ion beam as recorded by the ion detector in two operations. First, an intensity value is computed from the data using signal processing techniques [2]. Then this quantitative measurement is converted into a concentration based on the gain of the instrumental platform.

[2] Also called *chemometrics* approaches.

In the following, we first describe methods to compute an intensity value from the LC/MS image in Section 4.5.1. The algorithms presented there are based on the linearity between the concentration of a peptide and its computed intensity-value. Elaborate methods have been proposed to also take into account the multiple signals generated from the same peptide, and improve the quantification of low-intensity signals.

For converting the intensity value back to a concentration, the main difficulty is that the ionization efficiency in the ion source is intricately dependent on the chemical properties of each peptide. At the moment, the conversion factor must be estimated for each peptide, as is described in absolute quantification (Section 4.5.2). However, in relative quantification (Section 4.5.3), only the relative concentration of proteins in two (or more) samples are measured. Less experimental variation are observed in the latter case.

4.5.1 Computing an intensity value

A given protein generates a signal that is distributed over many pixels in the LC/MS image:

- The protein is digested into peptides as indicated in Chapter 2, so many peptide signals at different retention times and with different masses should be attributed to the protein.
- A given peptide with mass m is ionized to different charge states, which appear with m/z ratios m/z where z is an integer, usually between 1 and 4.
- A given ion is present in several isotopic forms with m/z differences corresponding to the additional neutrons.
- Finally, the signal of a homogeneous set of molecules, hereafter called a *peak*, is distributed over many pixels in the image because of the limited resolution of the chromatographic column and the mass analyzer.

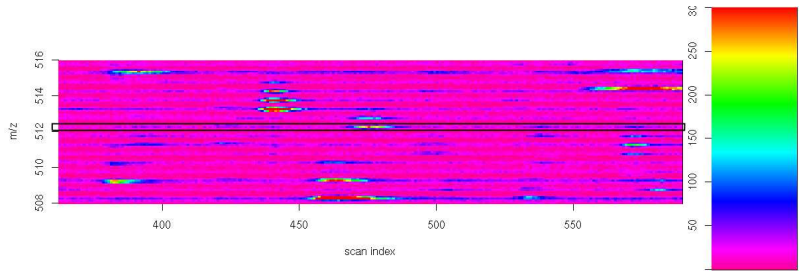

Figure 4.7: The simplest quantification method is to compute the integral of a single peak in a window around the signal.

The crudest method to obtain an intensity value for a peptide is to compute the area under the curve of a peak (see Figure 4.8); this corresponds to the integral of the intensity function. Although very simple, this method is widely used because experimental errors — as opposed to signal processing errors — are usually the limiting factor in proteomics experiments.

Indirect quantification methods have been proposed based on feature detection by template matching. After optimizing the template parameters, the area under the curve is deduced from the template shape rather than the sum of the ion intensities [LGR$^+$06, NF07, GMG$^+$99, JPBF$^+$04]. For example, there are five degrees of liberty to optimize the match between a Gaussian template and the LC/MS image: area under the curve, retention time, m/z and standard deviations in the two dimensions.

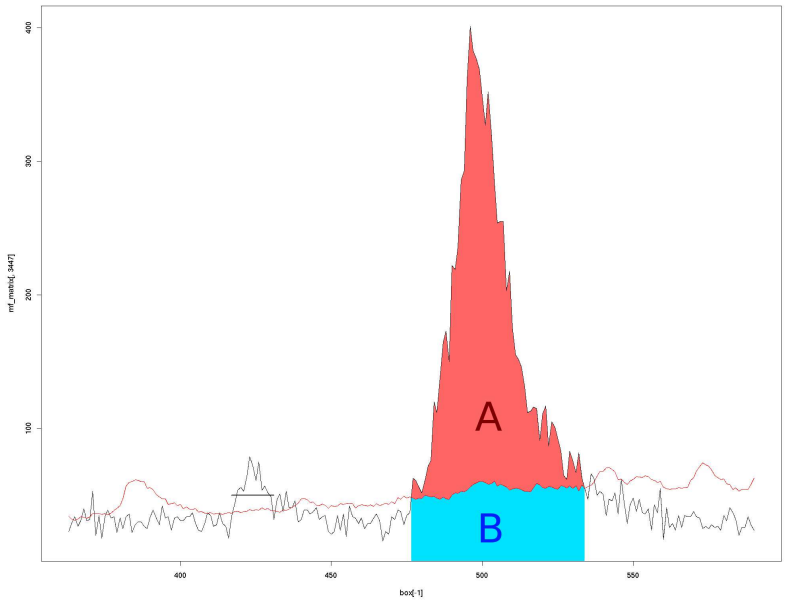

Figure 4.8: Area under the curve quantification in LC/MS. **A** corresponds to the intensity of the peptide signal. **B** corresponds to the contribution of background noise as estimated in Chapter 6.

In the previous methods, only one peak per peptide is used for quantification. To leverage additional information contained in the other peaks, several groups have proposed to combine the templates for each individual peak into a one-dimensional profile similar to a mass spectrum [WTF+07, STHG+07]. To construct such a profile, we need to know the relative positions of the peaks and their relative intensity.

Theoretical profiles can be constructed for isotopic motifs. Isotopic peaks are located at the same retention time [3] and at m/z ratios $m_0/z + i/z$ where m_0 is the mono-isotopic mass, z is the charge and i is the number of additional neutrons. The relative intensities in an isotopic profile can be computed from the chemical formula using a tool such as IPC (Isotope Pattern Calculator `http://isotopatcalc.sourceforge.net/`).
In proteomics studies, the relative intensities can be approximated based on the composition of a theoretical amino acid called averagine [SBM95], or other similar models [BHWW00, VJB08]. Theoretical profiles are the basis for deisotoping and charge-state deconvolution methods (see [LNRE05] for a review).

Unlike the distribution of signals among isotopic peaks, the distribution of ions among the possible charge states is difficult to predict. Although the different peptides from a given protein are present in the same concentration after digestion, they do not produce signals with the same intensity; this is due to different ionization efficiencies for each peptide. Consequently, theoretical models do not go beyond isotopic patterns.

Empirical two-dimensional profiles can be estimated from calibration experiments, by using methods from factorial analysis such as PCA [EGW87], N-PLS [Bro96] or PARAFAC [Bro97]. For more information on these methods and their statistical foundation, we refer the reader to [LMP00, KHO04]. The intensity measure is obtained after matching the profile to the image.

[3] because isotopes have nearly identical chemical properties

In his PhD thesis [Str08], Gregory Strubel presents a procedure for computing the intensity value based on two-dimensional profiles. The profile of a protein is a sum of Gaussian peaks. For each protein, calibration experiments are used to compute the parameters of the profile, i.e. the retention time and m/z position of the peaks, as well as the relative distribution of intensity among the different peaks.

To match the computed profiles to the LC/MS image, Gregory Strubel considers a Bayesian procedure, and selects prior distributions for the protein concentrations, the conversion factor for each protein, variations of the retention time and additive noise. A Gibbs sampler is used to simulate the distributions a posteriori and compute the mean value.

This refined algorithm is applied in the context of directed analyses, where the list of proteins of interest is known, and careful calibration experiments are available to compute the profiles. In the context of our thesis, the analysis is not directed, and we attempt to identify and quantify all the components of the biological sample.

4.5.2 Absolute quantification

Absolute quantification is only possible with an estimate of the conversion factor between the recorded intensity and the concentration of the protein. This factor is dependent on the sample preparation protocol, but may also depend on the concentration of other proteins in the sample (competitive ionization). For this reason, precisely calibrated absolute quantification is not a large-scale method, and is only available in targeted analyses.

4.5.2.1 Calibration curves

To build calibration curves, the intensity of the protein signal is measured in controlled experiments, with known concentration of the protein. The conversion factor is measured as the slope of the linear relationship between concentration and intensity. After some optimization of the experimental protocol, it seems relatively easy to obtain linear calibration curves.

This approach is not high-throughput because the calibration curves are peptide specific, and also depend on the sample preparation protocol. Moreover, it is necessary to produce the peptides of interest in known concentration to measure the calibration curves. Finally, calibration curves obtained for purified proteins may not be applicable when the protein is in a complex biological sample.

4.5.2.2 Internal standards

To remove some of the experimental variations, proteins may be quantified relative to the known concentration of an internal standard. Internal standards are molecules introduced in the sample in controlled concentration, and which have the same chemical properties — retention time and ionization efficiency — as a protein of interest[4]. However, the molecular weight of the internal standard must be different from that of the measured protein. Only isotopes match these requirements. In the case of tryptic digests, [BDA+07] proposed to synthesize isotopes of the whole protein instead of its individual peptides.

[4] As ionization efficiency depends on chemical properties, a given internal standard is only adequate for quantification of one protein.

Methods based on internal standards are low throughput methods like those based on calibration curves. Although internal standards can be used on different experimental contexts, they still need to be synthesized and are limited to the quantification of one protein. Additionally, only a few internal standards can be introduced in a biological sample without altering its properties. For instance, standards are likely to hide other small intensity signals from low-abundance proteins.

4.5.2.3 SRM and MRM quantification

Instead of measuring the intensity of a protein as the area under the curve in mass spectra, single reaction monitoring (SRM) and multiple reaction monitoring (MRM) methods consider the intensity of fragment ions measured in MS/MS spectra. Quantification performance is improved because this two-stage filtering approach focuses on the peptide of interest and removes a large part of chemical noise found in MS spectra. For more details, we refer the reader to [YC09, YBV08].

4.5.3 Relative quantification

Relative quantification is more precise because a lot of sources of experimental variation are removed by mixing the experimental samples, and applying sample preparation only once. In most methods, only two samples can be mixed, which limits the range of applications in biology. As the protein signal intensity may originate from one sample or the other, the biological samples are modified to introduce detectable differences in the LC/MS data.

Modification of the samples is usually done with isotope labels; one sample is tagged with a light label, and the other one is tagged with the heavy label. All proteins tagged with the heavy label have a fixed additional molecular weight δm with respect to their light counterpart. In LC/MS images, labelled proteins appear as pairs of signals with the same retention time[5] but a $\delta m/z$ shift on the m/z axis. δm must be large enough so that isotopic patterns of the two protein species do not overlap; $\delta m = 4$ is usually sufficient.

Methods in relative quantification can be classified depending on the stage of the sample preparation where the samples are mixed for comparison:

- metabolic labelling modifies the cells or tissues. The harvested samples are first labelled, mixed, then prepared according to the procedure described in Chapter 2.
- chemical labelling modifies the proteins or digested peptides after denaturation.
- label-free methods do not mix or modify the samples.

4.5.3.1 Metabolic labeling

The earliest point at which stable isotope signatures can be introduced in a sample is during cell growth or division. This involves changing the culture medium, for example by replacing light nitrogen atoms ^{14}N with the heavier ^{15}N atoms[6]. It is more practical to use stable isotope labeling by amino acids in cell culture (SILAC approach) where only a few amino acids are replaced instead of all molecules. Current methods use ^{13}C-arginine and ^{13}C-lysine, which ensures that all tryptic peptides carry at least one labeled amino acid.

[5] Isotopes have nearly identical chemical properties.

[6] In the ^{n}Z notation, n corresponds to the number of neutrons in the atom nucleus. The standard form of nitrogen is ^{14}N, the standard form of carbon is ^{12}C, and that of oxygen is ^{16}O. Additional neutrons weigh roughly 1 Da.

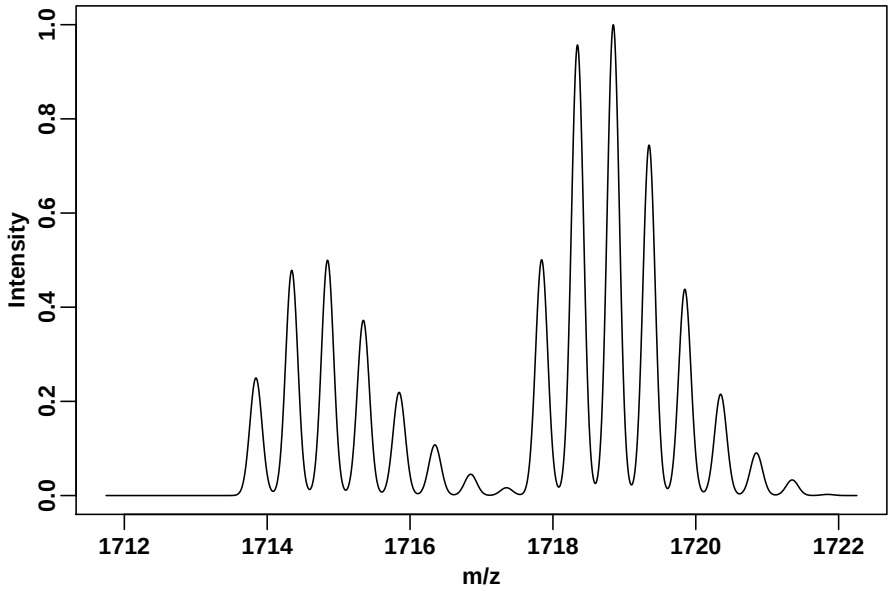

Figure 4.9: Simulated mass spectrum showing relative quantification with isotopic labeling. The mass spectrum contains two isotopic distributions corresponding to the light (+0 Da) and heavy (+8 Da) versions of insulin. In this example, the ions are doubly-charged, and the m/z difference is 4Da.

As the incorporation of the labels and the mixing of the samples precede sample preparation, metabolic labeling ensures that all sources of quantification errors in the sample preparation are excluded. Although metabolic labeling of higher organisms such as worms (C. elegans), flies (Drosophila melanogaster), rats or plants have been reported, this is not a routine procedure for anything other than cells.

4.5.3.2 Chemical labeling

Chemical labeling of proteins and tryptic peptides is carried out in vitro. Several types of isotopic labels can be introduced.

^{18}O **labeling** incorporates heavy oxygen at the COOH terminus of the peptides during digestion by trypsin (see Section 2.2). Digestion naturally involves replacing the oxygen atoms from the COOH function with oxygen from the water molecules in the solvent, and it suffices to perform the digestion in heavy-oxygen water.

Isotope-coded affinity tag (ICAT) labeling adds a label to the cysteine amino acids in the peptide sequence. The first versions of the tags used either zero or eight deuterium atoms for a 8 Da mass difference, but lead to retention time differences between the light and heavy proteins. This has been corrected in newer versions. A major drawback of ICAT labeling is that cysteine is a rare amino acid, and only a fraction of the peptide mixture can be labeled.

ITRAQ labels are special because the different tags have the same molecular mass. Distinction between the samples is achieved in fragmentation spectra because fragmentation occurs at different sites. Isobaric tags like ITRAQ do not increase the number of signals in LC/MS images, and can be used to compare several samples, up to eight in the current implementation.

4.5.3.3 Label-free methods

Although isotopic labels provide the best quantification performance today, they have several shortcomings. One is that the biological samples are modified, with sometimes limited incorporation of the isotopic labels; this requires more elaborate algorithms. Another problem is that the samples need to be available at the same time for mixing. Label-free methods alleviate these problems, but in turn need reproducible sample preparation and software preprocessing.

Currently, two label-free quantification strategies can be distinguished:

- measuring and comparing the intensity in LC/MS images of different samples
- counting and comparing the number of MS/MS spectra in LC/MS/MS analyses.

Comparing the intensity values between LC/MS images is easy to interpret. However, its implementation in practice is dependent on many signal processing steps that deal with the reproducibility of

- sample preparation (normalization problems discussed in Section 6.8),
- chromatographic separation (retention time reproducibility discussed in Section 6.2),
- the linearity between protein concentration and measured intensity (influenced by the baseline, cf Section 6.7, and competitive ionization effects, cf Section 3.2).

Reproducibility can be improved by introducing internal standards as presented in Section 4.5.2.

The spectral counting approach [LSYr04] is based on the observation that abundant proteins are more likely to be selected for MS/MS fragmentation and identification. Relative quantification of large changes have been demonstrated with spectral counting, but subtle changes require many MS/MS spectra from the same protein species. Spectral counting benefits from extensive MS/MS acquisition protocols, both for protein identification and for quantification, whereas the standard method trades MS/MS-based protein identification with the sampling rate of the peptide signals.

Summary

The main ingredient for quantification is the intensity and in particular the area under the curve of the peptide signal. Limiting factors are not in the computation of this integral, but rather in the low reproducibility of the analytical procedures.

When comparing two biological samples, technical variations can be overcome by labeling the samples with stable isotopes, mixing them, and performing LC/MS analysis of the mixture. As this is quite restrictive for biological applications, label-free methods are important, although they provide poorer quantification results. A very detailed review of quantification procedures in LC/MS can be found in [BSS+07] and in particular Table 4.1 as reproduced below.

	Application	Accuracy	Proteome coverage	Dynamic range
Metabolic protein labeling	Cell cultures only	+++	++	1–2 logs
Chemical peptide labeling (MS)		++	++	2 logs
Chemical peptide labeling (MS/MS)	Comparison up to 8 states	++	++	2 logs
Enzymatic labeling (MS)		++	++	1–2 logs
Spiked peptides		++	+	2 logs
Label free (ion intensity)	Targeted analysis of few proteins	+	+++	2–3 logs
Label free (spectrum counting)	MS/MS analyses	+	+++	2–3 logs

Table 4.1: Quantification methods for LC/MS analyses (adapted from [BSS+07]).

4.6 Conclusion

In LC/MS data, the position of a signal (its retention time but especially its m/z ratio) is used for identification, and the signal intensity is used for quantification.

Current instrumentation is neither accurate nor precise enough to identify peptides or proteins based solely on their m/z ratio. This information is supplemented by the m/z value of (a) tryptic peptides in Peptide Mass Fingerprinting, (b) fragment ions in MS/MS identification or (c) the retention time in aligned LC/MS images. The preferred method is MS/MS identification, but it can only identify a fraction of the signals in the LC/MS image.

For quantification, the label-free comparison of the area under the curve of the peptide signals is not a precise method because of technical variations in the sample preparation protocol. By mixing the samples to be compared, and after stable isotope labeling, accurate relative quantification results can be obtained at the cost of additional sample handling procedures and reduced coverage of the proteome. Some of the technical sources of variations can be corrected, as presented in Chapter 6.

Although MS/MS-based methods for identification and quantification (MRM) provide the best results, these methods are limited in terms of throughput. Fragmentation takes a lot of instrument time and there is a demand for methods without it. MRM quantification is necessarily focused on a (known) subset of peptide signals.

Chapter 5

Biomarker discovery

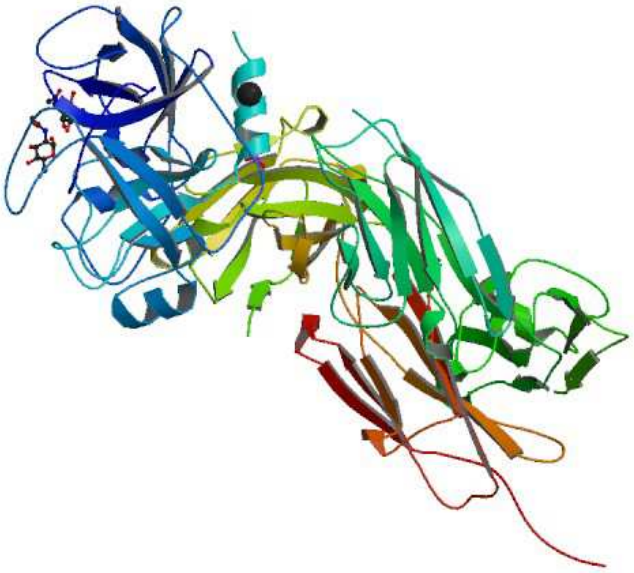

Figure 5.1: Three-dimensional representation of PSA (Prostate Specific Antigen).

5.1 Protein biomarkers

5.1.1 The biomarker concept

Originally, a *biomarker* is a substance found in the blood, urine, or the tissues of a living organism that is detected in elevated amounts in patients with a certain disease [LFP03]. This notion of what a biomarker is has since been extended to biomarkers that are in lower-than-normal abundance and for other applications like disease classification, and monitoring the disease progression. In statistics, a biomarker corresponds to an explanatory variable.

Several types of molecules can be used as biomarkers. Genetic biomarkers have lead to simple and efficient tests for genetic diseases. Microarray technology can be used to detect RNA

biomarkers. However, there is little correlation between the abundance of RNA and the quantity of the corresponding proteins. As proteins are responsible for most cellular functions, it is believed that the protein concentration can better indicate the state of the biological system under consideration.

5.1.2 Biomarker detection for clinical applications

The notion of biomarkers is used mostly in a clinical setting. Biomarkers are particularly useful when they can be detected in samples that can be collected using non-invasive procedures, without hazard to the patient's health. Such samples comprise blood, serum and urine and other body fluids. Tissue samples obtained in biopsy are rather used for diagnostic confirmation because biopsy requires surgical procedures.

Blood, serum and urine are complex mixtures. Biomarkers are not expected to be prominent in those samples; LC/MS technologies are promising for detecting biomarkers due to their sensitivity, i.e. their ability to detect, identify and measure the concentration of low-abundance proteins in spite of the sample complexity. Even if the disease does not target serum, this fluid permeates all the tissues in the human body, and traces of the activity of all cells are released in serum. These are the potential biomarkers.

Because of large natural variations in the biology of individuals (differences in diet, genetic background, lifestyle and so on), potential biomarkers need to be validated on large numbers of samples. Consequently, a suitable technology should also be affordable, quick (high throughput) and reproducible.

5.1.3 Biomarker discovery with LC/MS

Biomarker discovery is the task of finding new molecules (proteins) with predictive power for disease diagnostic, classification or monitoring. LC/MS addresses the requirements for biomarker discovery in a clinical setting:

- complete proteomes can be analyzed (exhaustive search space),
- sensitivity, the technology can detect and quantify low-abundance proteins,
- high-throughput because LC/MS experiments take on the order of an hour depending on the LC separation[1], and identifies and quantifies thousands of proteins,
- multiple molecular signals can be combined for increased sensitivity and discriminative power.

Single biomarkers are proteins that indicate by themselves the presence of a disease. For cancer diagnostics, some single biomarkers for cancer are known such as Prostate Specific Antigen (PSA, marker for prostate cancer) and alpha-fetoprotein (AFP, testicular cancer and ovarian cancer). However, measurements of tumor marker concentrations are not sufficient for diagnosis for the following reasons:

- marker levels can be elevated under benign conditions,
- marker level may not be elevated in early stages of the disease,
- many markers are not specific to a particular type of cancer,
- some tumor lead to elevated levels of most proteins, which can be misinterpreted as a normalization problem (cf Chapter 6).

Currently, the main application of tumor markers is for assessing the response to treatment and to monitor recurrence (e.g. CA19-9 for monitoring pancreatic cancer).

[1]This does not include the sample preparation phase which may take a few hours. Most of this time corresponds to tryptic digestion.

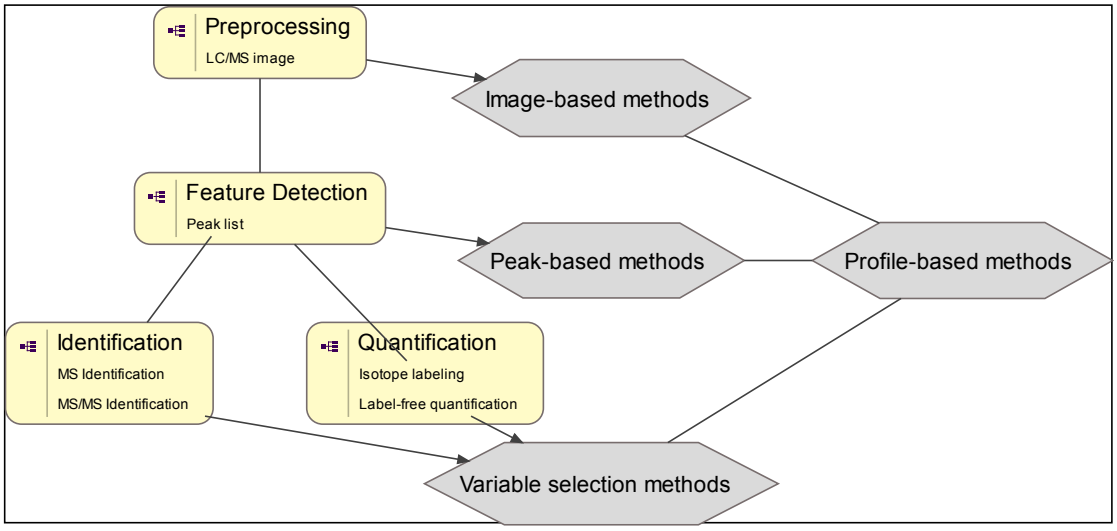

Figure 5.2: Classification of biomarker discovery methods based on the type of input data. Profile-based methods can be proposed as extensions of the other methods.

Biomarker discovery in LC/MS experiments combines identification of proteins and quantification. The basic approach is to find proteins with different concentration levels in the patients and the controls. When supplied with the list of proteins in the sample and their concentration, the whole range of machine learning methods are available. Most of the research in the field has focused on supervised classification algorithms as reviewed in [SM06], including discriminant analysis, decision trees, neural networks, support vector machines etc.

Current approaches can be classified into three types:

1. Variable selection methods find the proteins that best classify the samples based on protein intensity. These methods are not specific to LC/MS data, use the results of identification and quantification but do not account for the mechanism of acquisition of LC/MS images. A review of these methods can be found in [SM06, HB06].
2. Profile-based methods, first proposed in [PAH+02], try to detect biomarkers as proteomic profiles or patterns. These can combine several biomarkers for classification. A nice review of these methods can be found in [PI06].
3. Alignment-based methods (peak-based or image-based) compare signal intensity between experiments after correcting the retention times in LC/MS images. We have reviewed these methods in [VLTTK+08], and reproduce the text in Section 5.2.

5.2 Discovery of potential biomarkers from aligned sets of LC-MS experiments

Recent developments in algorithms for processing LC-MS images have expanded the application of LC-MS to many fundamental biological questions, such as the discovery of differences in protein profiles between two or several classes of biologically distinct samples; e.g. healthy and disease groups. In particular, alignment algorithms, as reviewed in Section 6.4, are making it possible to supplant complex labeling approaches with a simpler, label-free, approach that relies on the direct comparison of aligned images between multiple LC-MS runs.

In this Section, we focus on the labelfree approach and its potential to speed up the discovery of both molecular markers for dissecting disease processes and novel targets for pharmaceutical

and/or diagnostics developments for complex human diseases. Since relative protein concentrations can be treated as unique and informative indicators of individual phenotypic state, proteomic biomarkers enable the classification of different disease subtypes, progression stages, and even individual responses to different treatment strategies, thereby allowing the development of systematic approaches to early disease prevention and treatment [Han03].

Despite its tremendous potential, several pressing challenges remain both in the development of the technological platforms for quantitative and comparative profiling of complex biological samples, such as tissues or body fluids [BMEC05], as well as in the development of statistical and computational methods for effective study design, data analysis and validation of the results [LE05, HCMB05, MAB+07].

In particular, some authors have questioned the biological relevance and reproducibility of the first cancer biomarker discoveries, based on SELDI-TOF or MALDI-TOF protein-profiling technologies from serum samples [Dia04a, Dia04b, BCM05, BMEC05]. It has been observed that the near-perfect classification results could largely be attributed to the presence of experimental artifacts that may severely corrupt the LC-MS images, which could be avoided if careful attention was paid to potential sources of experimental bias before the experiments are run [HCMB05]. Although LC-MS provides an additional dimension of separation, as compared with SELDI or MALDI techniques, similar considerations on study design and preprocessing methods are similarly critical when applying LC-MS-based protein-profiling approaches to differential analysis (as will be discussed in Section 5.2.3).

Two types of study objectives are commonly encountered in differential analysis of LC/MS experiments [SRDM03]:

1. class comparison aims to determine whether the protein profiles are different between the given classes of samples, (e.g. different stages of breast cancer), and if so, to identify the most discriminative LC-MS image features. This study objective can be addressed using a standard statistical hypothesis testing framework, while carefully controlling for the multitude of comparisons made [LE05].
2. In class prediction studies, the emphasis is on developing LC-MS image-based multivariate discrimination functions (or classifiers) that can accurately predict the true class membership of a new sample according to a set of key features such as detected peaks or image patterns (here referred to as potential biomarkers). For this type of problem, several machine learning-based classification techniques or their combinations have been applied to MS data [LE05, HB06, BAVL06, WZP+06].

In many cases, comparative LC-MS studies deal with both class comparison and prediction problems, both requiring that class labels of the biological samples are predefined beforehand, while the differences originate mainly from the type of measured features that are considered as potential biomarkers.

Although all data points in the aligned LC-MS images represent potential biomarkers, it is important to identify those key features that are discriminative and robust against both technical and biological variation present in the data. Such task, called feature selection is typically performed as a preprocessing step to limit the number of features, either using class comparison (filter approach) or prediction (wrapper approach). These "supervised" approaches are built on "class labels" that indicate the class for a certain set of samples.

Class comparison and class prediction do not necessarily lead to the same set of features, even when performed on a single dataset [PPH06]. An alternative approach is to use unsupervised techniques (clustering), which do not require any class labels, but use the multidimensional distribution of the dataset to assign the labels. Several data mining-based clustering techniques can be used both for sample discrimination [HB06], as well as for feature selection and summarization

[LE05, ACvG+06]. Depending on whether feature selection is performed on pre-processed intensity data or on the detected signal peaks, strategies to identify differences between two classes of samples can be divided into imagebased and peak-based approaches.

5.2.1 Peak-based approaches

In the conventional analysis workflow for LC-MS data, peak detection is a critical step. The subsequent analysis steps, including image alignment and biomarker discovery, are based on comparisons of the resulting experiment-specific lists of detected peaks, in terms of their numeric attributes like m/z location, elution time and peak height. Through peak detection, it is possible to reduce the amount of data from millions of image intensities to some hundreds or thousands of peak-specific features. This was considered crucial when computational power was not as available as today.

Another advantage of using peaks as basis for biomarker discovery is that a peak detection step can filter partially both technical measurement noise and confounding biological variation by focusing on clearly detectable peaks, which are more likely to represent defined proteins, protein fragments or peptides. Detecting entire isotope patterns instead of individual peaks may further improve the reproducibility of the results by reducing the number of detected false positive noise peaks [KO07]. Due to the challenges in peak detection, especially in low-resolution data, there have recently been a growing number of studies performing direct comparison of the pre-processed image intensities (reviewed in Section 5.2.2).

An essential prerequisite for systematic biomarker discovery using the label-free LC-MS-based approach is an accurate and reproducible quantification of differential expression through comparison of peak-features between aligned LC-MS runs. Utilizing an IT LC-MS instrument and a dynamic time-warping algorithm in their MASSVIEW software, Wang et al. [WZL+03] demonstrated a nearly linear relationship between the LC-MS peak area and the peptide ion concentration for a variety of compounds introduced into human serum, and a modest coefficient of variation across integrated peak intensities from independently processed samples.

Recently, Wang et al. [WWZ+06] showed that highly reproducible data can be obtained between replicate LC-MS runs, in terms of nearly ideal Pearson's correlation for both peak areas and elution times of identified peptides, when an IT or Fourier transform mass spectrometer is used. To demonstrate the applicability of their analysis software, QUOIL, for identification in differential experiments, they also showed that observed average ratios of peak areas relative to those in a reference sample correlated well with the known abundance ratios.

While the above studies did not contain any feature selection step, Silva et al. [SDD+05] carried out a detailed peakbased differential expression analysis that relies on detecting relative changes in experiment-specific lists of features called accurate mass-retention time (AMRT) components. Their software, Expression Informatics, first performs ion detection in m/z space by a maximum likelihood-based algorithm that performs de-isotoping, charge-state reduction, and peak quantification. It then detects the AMRT components at the elution time of maximal intensity, if their intensities exceed a minimum detection threshold. The threshold is defined manually by the user or automatically by the software.

Utilizing the relatively high mass precision of a Q-TOF instrument, and clustering-based matching of the AMRT component lists among multiple runs, Silva et al. [SDD+05] were able to determine differences in relative abundance of a small subset of proteins spiked into a complex protein background (human serum).

Finally, the software performs multiple pair wise comparisons among the matched AMRT components between different sample classes and assesses the statistical significance of the observed intensity ratios with the Student's t-test. However, neither multiple comparison adjustments for the significance levels nor a predictive discrimination between physiologically distinct phenotypes were performed.

In another study using high-precision instrumentation, America et al. [ACvG+06] applied a peak-based differential analysis to identify discriminative protein patterns between tomatoes at two distinct stages of fruit ripening. Highly complex samples were analyzed by nano-LC coupled to a Q-TOF MS instrument, using three technical replicates to assess its reproducibility.

Their specific software tool, MetAlign, first performs data denoising (filtering) and reduction (binning), followed by peak detection and alignment, and then compares peak intensity values of the replicate LC-MS images between the two classes of samples using a pairwise t-test.

To reduce the dimensionality of the feature space, originally composed of the intensity values of all significant peaks, the normalized peak lists are further analyzed with a principal component analysis (PCA). PCA projection using the first two principal components could already reveal a clear separation between the two classes of tomato samples, while the triplicate runs were mapped close to each other. Although these results are very promising, they do not demonstrate whether the method can additionally predict the class of an unknown sample.

Radulovic et al. [RJR+04] developed a comprehensive software suite of algorithms and data mining methods to support large-scale LC-MS-based profiling and biomarker discovery in complex protein mixtures using common experimental procedures and MS instrumentation. For data preprocessing, it contains signal filtering and amplitude normalization as well as peak detection, quantification, binning, and alignment steps.

However, no machine learning or data mining algorithms were used in the final biomarker discovery step, but the discrimination power gained from these advanced image-processing algorithms was evaluated by simply visualizing the peak overlaps between aligned pairs of LC-MS images. Moreover, the limited number of liver samples extracted from two physiologically distinct classes of inbred mice that were used in the evaluation (two samples per class for training and only one for testing its predictions) hinders any concrete conclusions on the predictive power and hence applicability of these algorithms to practical biomarker discovery studies. Another potential limitation may originate from binning of the detected peaks according to their nominal masses, which may result in discretization artifacts and significant loss of relevant information.

5.2.2 Image-based approaches

An inherent disadvantage of the peak-based methods is that detection of peaks in the LC-MS images is prone to errors, which can complicate the image alignment process and lead to spurious biomarkers. Accordingly, many groups have proposed an image-based differential analysis approach that eliminates peak detection and performs the comparisons directly on the pre-processed LC-MS image intensities.

When the image comparison is based on the original intensity values, numerous additional features such as peak shape or elution range – which are lost in simple peak detection – are available for correcting technical distortions in the image alignment phase and for revealing discriminative features in the biomarker discovery phase.

With the image-based approach, however, potentially strong correlation between the image features as well as technical noise blurring m/z values, and the fact that peptides do not elute instantaneously should be carefully considered [LE05]. Such a direct approach is also computationally rather demanding as the intention is to shift complexity from the peak detection to the image comparison phase, and it was not until recently when computational power became available enough to allow direct image-based differential analysis methods.

Based on their alignment algorithm, CHAMS, Prakash et al. [PMW+06] carry out the comparison of LC-MS images across multiple experiments directly on the image level without data reduction steps such as peak detection. The alignment results can be conveniently visualized in terms of the scoring function that quantifies the proportion of similar pairs of m/z and intensity values between the runs.

Then, a feature detection algorithm is used on the aligned images to identify those image patterns that are discriminative or common across the different experiments. As features, Prakash et al. [PMW+06] consider the measured intensity values, provided that they are large enough and contained other large intensity values in the nearby spectra and within their isotopic range.

Common features are first searched from the aligned replicate runs under the same condition, and discriminative features are then identified from the aligned images originating from the different conditions. The intuition behind this approach is that biomarker discovery on the basis of multiple experiments will be more sensitive and specific than by identifying features first and then comparing these across multiple experiments.

Preliminary evidence supporting the benefits of the algorithm in biomarker discovery is demonstrated on data from several instruments, ranging from LCQ to Q-TOF and LTQ. However, Prakash et al. do not perform comparative evaluations in terms of either class comparison or class prediction performance. Moreover, the proposed algorithm was evaluated only visually in a rather simple experimental setting, in which one spiked-in peptide was added into a background mixture consisting of only four proteins.

Wiener et al. [WSDY04] introduced an automated differential mass spectrometry method (dMS) for detecting significant differences between multiple images obtained using a labelfree LC-MS approach. The method directly compares intensities at each elution time and m/z ratio, and outputs a ranked list of differences between samples under two conditions. No actual alignment was originally used but the measurements from different runs were made comparable by binning the intensities into a grid using equally sampled points for both m/z ratio and elution time.

Multiple LC-MS runs were used for accounting the variability in the measured intensities, thus enabling detection of subtle but consistent differences also in low-abundance peptides, while ignoring large but irreproducible differences in peptides present at large concentrations. Statistical significance of the difference in total intensity at each bin between the two given conditions was assessed with a t-test. Technical variation that may confound the true differences was reduced by considering those intensity changes only whose statistical significance persists in time intervals longer than a given time limit.

Good sensitivity and specificity of the dMS method was originally demonstrated in a controlled experiment involving several spike-in mixtures with known peptide differences analyzed with an IT instrument [WSDY04], and recently a modified version was re-evaluated on similar experiment using a Fourier transform instrument [MWS+07].

SpecArray, by Li et al. [LYK+05], is a software suite that takes a set of LC-MS data as input and generates a peptide versus sample array, storing the relative abundance of peptide features matched in all samples. A peptide extraction algorithm defines peptide features by considering the peptide's monoisotopic mass, charge and elution time and uses MS signals to determine the abundance and the S/N of the feature. Peptide features occurring in repeat LC-MS analyses are combined to yield unique peptides, possibly comprising different charge states.

A pattern-matching algorithm is applied to align peptide features between different samples (based on their charge, mass and calibrated elution time), by evaluating an elution-time calibration curve between any two samples, in order to correct any elution time shifts between two LCMS analyses. Information on discriminatory peptides is collected, only after which targeted MS/MS is used to identify corresponding proteins.

Identification of discriminatory peptide features between samples of different characters is done by subjecting the peptide array to unsupervised clustering analysis (hierarchical or k-means) or to discriminant analysis (Student's t-test or linear discriminant function), and by consequently evaluating the p-value of the discriminatory peptide features for relative abundance in the peptide array.

In an experiment aimed to profile serum proteins of five male and five female mice, peptide features were discovered that distinguished male mice from female mice, the difference in proteins of which appear to be insignificant from a statistical point of view.

Listgarten et al. recently conducted a very comprehensive LC-MS-based computational biomarker discovery study [LNR+07]. Their HMM-based alignment approach uses a continuous profile model that performs the elution time alignment and amplitude normalization simultaneously.

After the preprocessing phase, a 2-D Hamming filter is applied separately to each of the binned intensity matrices to correct for small local distortions in the aligned image intensities, originating, e.g. from binning artifacts. Similar to Wiener et al. [WSDY04] they evaluate the intensity differences at each image bin between two classes of samples using a t-test. However, instead of using a defined minimum threshold on the time intervals, Listgarten et al. [LNR+07] apply once again a 2-D Hamming filter but this time on the matrix consisting of the bin wise t-statistic values to leverage the differences across both dimensions.

The method is tested on a controlled human serum spike-in experiment analyzed with an IT MS, showing that at the t-statistic level corresponding to recall of 50% and 50% precision could be obtained when using seven replicates. While the class comparison accuracy is relatively far from the optimal of 100%, they show that a perfect class prediction result, in terms of cross-validation, can be achieved using a regularized logistic regression classifier with features extracted from a 14-D PCA projection of the image intensities [LNR+07]. The utilization of all the intensity bins, however, limits the usefulness of the classifier as a practical diagnostic tool, which typically has to rely on one or a few biomolecules.

5.2.3 Challenges in computational biomarker discovery and validation

The above-mentioned initial investigations support the potential of the label-free LC-MS-based approach to protein biomarker discovery. Due to the small sample sizes, however, most of these studies could not thoroughly evaluate the predictive power of the algorithms but mostly demonstrated the reproducibility and discriminative power of the techniques using statistics like Pearson's correlation or Student's t-test.

While such evaluations are useful as initial proof-of-concept and, later, for reducing the number of potential biomarkers, the generalizability, and hence the true benefits for practical biomarker discovery studies, can be assessed only by re-evaluating the computational predictions on independent data sets [MAB+07, LW05].

In the absence of enough samples to allow dividing them into two distinct training and test groups, cross-validation can be used to limit the risk of over-fitting [SRDM03]. Even if LC-MS profiling is used merely as exploratory tool, it may still be useful to first assess the generalizability of the potential biomarkers computationally.

After the reproducibility of the measurements has been verified, confirming the reproducibility of the prediction accuracy on independent test samples should be considered as a minimal necessary condition for the validity of the classifier, and of the features, it is built on. If the reproducibility of these candidate biomarkers is preserved on independent samples as well, one may consider proceeding to the next phases of biomarker development, involving clinical immunoassays and longitudinal studies [LW05].

Assessing the generalizability of the discovered features using blinded and heterogeneous samples is especially important when using large numbers of LC-MS image-based features extracted from the observed intensity distributions.

While such morphological characteristics can facilitate the image alignment and biomarker discovery tasks, this extra information can sometimes be strongly misleading and irreproducible due to many sources of technical measurement variability, e.g. accidental co-elution of features or spurious features that present noise or background ions. Such artifacts can be easily overlooked when relying on automated feature finding methods [CBFC06]. However, proper validation of the results (e.g. by cross-validation) has the potential to eliminate such spurious image features.

In addition to the technical variability, there also exist many sources of biological variability, for which treatment is beyond the capacity of standard preprocessing techniques and cross-

validation studies if they are only performed on a set of similar samples. Even in the case of a single LC-MS dataset, the cross-validation prediction accuracies can be very similar, even though the discovered set of key features overlap only modestly, depending on the subset of individuals used for feature selection. This is because the global MS experiments do not only express the phenotypic differences of interest but also many other confounding factors attributed to differences within or between individuals that are related to other aspects of the physiological state than the actual phenotype of interest [HMF+05].

In many respects, the computational challenges in the MS-based biomarker discovery strategies are very similar to those already encountered when using gene expression profiling with high-throughput microarray [SRDM03]. In particular, most of the statistical issues suggested in the above-mentioned LC-MS-based differential analyses are standard in microarray data analysis [ACPS06]. These include, e.g. the observation that a simple fold change criterion is not as effective as the t-test statistic, that a non-parametric test may in some cases be preferable [MCK+05], or that outliers can seriously bias the statistical values, and that the effect of multiple testing should be corrected [WWZ+06].

Additional issues that also could prove to be useful in the analysis of global LC-MS data include: (i) the utilization of variance shrinkage in class comparison to improve the variance estimates when sample sizes are small (modified t-tests), (ii) estimation of false discovery rate (FDR) rather than controlling for family wise error rate (FWER), and (iii) the merits of different types of resampling techniques (e.g. bootstrap or permutation) in non-parametric tests or in test statistic selection [ACPS06, EFLA07].

Perhaps the most critical step that affects all the downstream analyses is the experimental design, including the decision of the plausible sample size and the type of replicates used. First, standardized experiments can make it possible to combine data sets from different laboratories and thereby increase the statistical power. Secondly, while technical replicates allow estimation of the measurement reproducibility, biological replicates are essential when the aim is to make general inferences about populations through sample data, as is the case in biomarker discovery.

Since label-free LC-MS-based profiling is an active area of research and development, the relative strengths and weaknesses of this approach to protein biomarker discovery have not yet been fully elucidated, and it is likely that new and improved methods will be introduced in the near future. While there is a fundamental lack of any systematic comparative studies, the image-based approach appears attractive because it operates on the complete intensity data and therefore does not lose any information valuable for image alignment and biomarker discovery. Accordingly, warping and matching images directly, rather than developing yet more elaborate peak detection techniques might be a promising direction.

While most current methods assume that the image features are directly comparable after the elution-time alignment step, compensating for variation in both dimensions could further facilitate the matching between the corresponding peptide signals, similarly to the alignment of 2-D gels [DDY03, ASNN05].

Well-designed LC-MS experiments and comparative studies will be needed to fully elucidate the real benefits of the different methods, relative to those from more traditional protein biomarker discovery approaches based on 2-D gels, isotope labeling, or MALDI technology [WAF+03]. Traditionally, different researchers have applied varying validation methods to non-standardized and rather small data sets to evaluate their customized algorithms and software solutions. Freely available reference datasets covering multiple instrumentation platforms, that are becoming available [KEH+08], allow direct comparisons of relative merits of different approaches that are likely to be closely linked to the precision of the MS instrument being used [LNR+07]. For comparing biomarker discovery approaches, however, public datasets with a sufficient number of biological replicates are critically needed.

After candidate features have been detected and validated, they can be analyzed at the MS/MS level or using identification approaches reviewed in Section 4.4. Identification of the potential biomarkers is critical for understanding the mechanisms underlying disease processes and if the panel of key biomarkers is to form the basis of a simplified and widely adopted diagnostic test.

Due to the limitations of the present technology to detect the subtle changes in the abundance of many important biomarkers from complex samples, like bodily fluids, it has been argued that differential analyses should be performed on a selected set of disease-related candidate proteins only [AAC+05]. The potential limitation of such targeted approach is the possibility of biased selection and limited usability when dealing with less-studied pathological conditions. The requirement of using only peptides with reliable identification in a particular experiment can also be problematic since the MS/MS experiments are relatively time-consuming and incomplete for complex mixtures, which increases the risk of missing several protein features of interest and therefore makes global comparisons extremely limited.

Image-based approaches can avoid these problems by first identifying all discriminative and reproducible image features from the multiple LC-MS runs, without relying on any prior information on the proteins of interest or selected peak model. In a second step, only those image regions could then be subjected to targeted MS/MS analyses for revealing protein evidence. This data-driven approach to biomarker discovery may not only save computation time but has also the potential of providing more sensitive results than traditional proteomic approaches.

5.3 Summary

LC/MS is a technology with a lot of potential for biomarker discovery because it combines sensitive detection with large-scale coverage of the proteome, identification of the interesting proteins and quantitative information. It also only requires little amount of biological material.

Biomarker discovery corresponds to finding explanatory variables, often in a supervised classification context. From the point of view of machine learning and statistics, this is a daunting task because individuals are few and there are thousands of observed variables. A related problem is to find differentially expressed genes from microarray data.

Standard machine learning approaches can be applied to the list of proteins and their quantification information. However, these methods ignore the processes involved in identifying and quantifying the signals and fail to detect some of the biases. One current research area uses the intensity measurements directly, sometimes without identifying the signals, as reviewed in Section 5.2. Another is to find groups of biomarkers or proteomic profiles that combine information from many protein indicators for increased sensitivity and specificity.

Chapter 6

LC/MS images and preprocessing

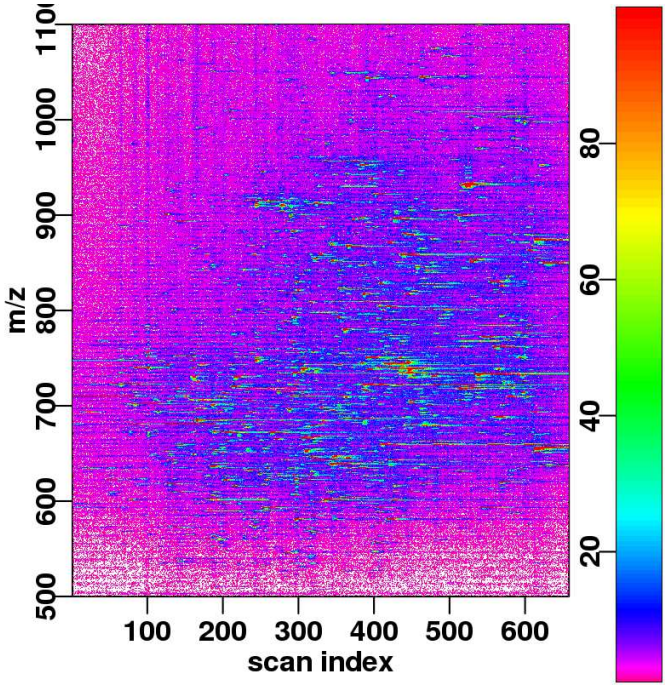

Figure 6.1: An LC/MS image

A LC/MS platform is an instrument of great precision, but it is also very sensitive to experimental conditions. Consequently, measurements acquired on a LC/MS platform are subject to many distortions that need to be corrected before subsequent analyses. In this chapter, we present the types of data contained in LC/MS images and the associated preprocessing methods: retention time alignment, m/z calibration and intensity calibration.

Preprocessing algorithms affect all the later analyses carried out on the data. The following table summarizes the effects of preprocessing on protein identification, quantification and biomarker discovery. For example, intensity calibration affects quantification and biomarker discovery but not identification.

	Retention time	Mass-to-charge	Intensity
Identification	+	+	
Quantification	+		+
Biomarker discovery	+	(+)	+

In LC/MS-based proteomics, identification of the proteins in the sample is based on the retention time and m/z ratio of each signal, but hardly on its intensity. All identification methods use the m/z value as the primary information, and require accurate calibration of the m/z values. In image-based identification as described in Section 4.4, the retention time of a signal supplements its m/z value on low m/z accuracy instruments or in the case of complex proteomes.

Quantification of proteins is based on the intensity of the signal. Intensity calibration is used to remove the contribution of background noise to the signal intensity, and remove systematic variations in the sample preparation protocol. Retention time calibration is required for comparing pixel intensities in different LC/MS images for label-free quantification approaches (see Section 4.5).

Preprocessing affects biomarker discovery through the combination of identification and quantification. Methods based on the comparison of unidentified, raw image intensities as described in Section 5.2 depend directly on accurate retention time alignment and intensity calibration. Methods based on the comparison of lists of proteins rely on the results of identification and quantification algorithms and are affected indirectly.

In this chapter, we describe the LC/MS images to be analyzed and the preprocessing algorithms. We first give an overview of the type of data used to make LC/MS images as well as numerical details on typical and ideal experiments in Section 6.1. This gives an indication of the potential and of the limits of signal processing for LC/MS images. We then proceed to the calibration of each type of measurement acquired on LC/MS platforms:

- retention time correction is presented in Sections 6.2, 6.3, 6.4 and 6.5,
- m/z calibration is presented in Section 6.6,
- intensity calibration is presented via baseline removal in Section 6.7 and intensity normalization in Section 6.8.

6.1 LC/MS images

This section presents data acquired on a LC/MS platform. We describe the data type but also the content of LC/MS images in terms of signals, noise and contaminants. The examples provided insist on the fact that LC/MS images are quite different from natural images and that they require specific processing algorithms.

If not otherwise noted, the LC/MS images in this chapter were generated from the data set in [KEH+08]. This data set corresponds to the analysis of a standard protein mix of 18 proteins on different LC/MS platforms, and is publicly available for download at
`http://regis-web.systemsbiology.net/PublicDatasets/`

6.1.1 Type of data

Detection The ion detector in the mass spectrometer measures the intensity of the ion beam after separation in the mass analyzer. Depending on the type of detector, the intensity is reported as floating-point values or integer values. These are always positive numbers. In particular, time-to-digital apparatus used in TOF instruments usually report integer numbers which should

correspond to the number of ions reaching the detector. On FT-ICR instruments, the amplitude of the Fourier transform of the measured signal is reported as a positive number.

Mass spectra Mass spectra are histograms, i.e. they report the intensity of the ion beam as measured by the ion detector as a function of m/z value. Although scanning of the m/z range takes time, this is negligible compared to LC separation, and a single retention time is attributed to all the data points that belong to the same mass spectrum. Each mass spectrum is presented as a vertical column in the LC/MS image.

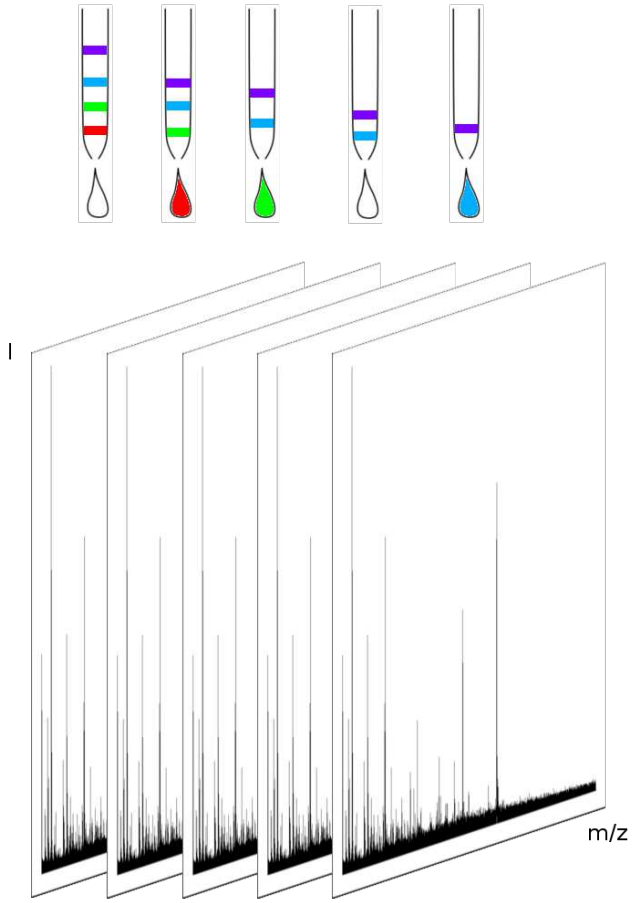

Figure 6.2: Coupling between LC separation and mass spectrometry. A mass spectrum is acquired for each droplet (fraction) eluting from the LC column.

LC/MS images When mass spectrometry is coupled to LC separation, the mass spectrometer records one mass spectrum for each fraction. In electrospray ionization, the retention time corresponds to the time point at which the sample droplet is transferred from the tip of the column to the mass analyzer. In MALDI ionization, the coupling is done off-line: the droplet is first deposited on a metal plate and mass analysis can be performed at later times.

Mass spectra are naturally ordered on a retention time axis. Following on this idea, mass spectra are usually assembled into LC/MS images such that each mass spectrum corresponds to a column in the image. In most setups, signal from a peptide species is distributed over several

consecutive fractions, and thus consecutive mass spectra. This elution profile is the basis for the feature detection method presented in Chapter 8. In the figures, the horizontal axis is not retention time but mass spectrum index in the data set because MS spectra are not acquired at evenly spaced time points.

MS/MS spectra Fragmentation spectra are ordinary mass spectra of a particular sample as described in Chapter 4. They share the same characteristics as their MS counterpart. On the order of three to five MS/MS spectra are interleaved between each MS spectrum in standard protocols. Few protocols acquire more than a dozen MS/MS spectra for each MS spectrum because liquid chromatography cannot be interrupted and sample is lost while fragmentation takes place.

Conclusion Data acquired on a LC/MS platform are triplets (retention time, m/z ratio, intensity) that are usually represented as images to enhance the continuity in the LC separation. In the following, we will represent the data as a function $\mathcal{I}(t, m)$ on the two-dimensional plane (retention time, m/z value).

6.1.2 File format and software tools

LC/MS proteomic analysis is a newborn field, and there are still many incompatible file formats. Each mass spectrometer generates data in a proprietary file format according to the vendor specifications, e.g. `.raw`, `.wiff`, `.t2d`, etc. Using the native file formats is subject to licensing fees and proprietary libraries that only run under the Windows platform.

To improve sharing of the data files as well as analysis with alternative software, the scientific community has proposed open file formats:

- `mzXML`, as described in [PEH+04]
- `mzData` from HUPO (Human Proteome Organization[1])

These two file formats are converging into `mzML` [2], for which the specification has been published in June 2008, and is still being updated.

Most proprietary software do not support the open formats. While there are converters available, the conversion process is still sketchy and requires both experience and access to the computers on which the mass spectrometer drivers are installed.

In this thesis, we use the `mzXML` file format. Most of the analyses were carried out using R, but we also developed tools in OCaml and C to solve some technical problems: segmentation faults, memory and speed requirements, etc. These include an `mzXML` parser for file input as well as tools for the visualization of LC/MS images.

6.1.3 Information in the data

Data complexity According to [NMA+05], bacteria like D. radiodurans have around 3,000 proteins, which are split into 125,000 peptides by tryptic digestion in the sample preparation phase. In the case of yeast whole cell lysates, 6,300 proteins and 416,000 peptides can be expected. There are 41,000 human proteins and 1.7 million associated peptides. These peptides are present in the LC/MS images generated from serum samples for biomarker discovery. Fortunately or not, most signals are not visible in the data because they are below noise level. Indeed, proteins exist in a wide range of concentrations. It is estimated that there is a factor of 10^6 or maybe 10^{10} between the concentration of common and low abundance proteins [DA06].

[1] `http://www.psidev.info/index.php?q=node/80#mzdata`
[2] `http://www.psidev.info/index.php?q=node/257`

Signals With extensive numbers of pre-fractionation steps, current LC/MS analyses are able to identify up to 10,000 proteins in a biological sample. In a single LC/MS image, identification of 1,000 to 2,000 proteins is possible.

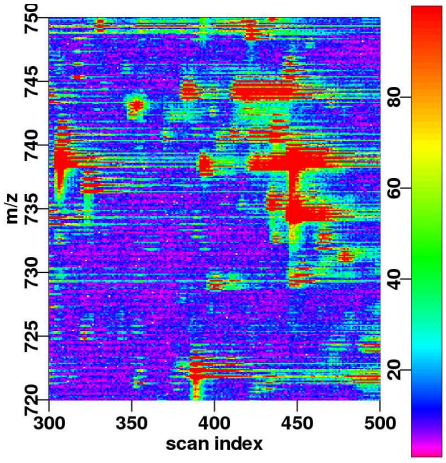

Figure 6.3: Peptide signals.

Molecular species corresponding to the same protein or peptide are naturally present in the LC/MS image at a different m/z ratios. Each peptide may have several charge states, and a varying number of additional neutrons. Isotopic ions with the same charge z form *isotopic patterns*: horizontal segments spaced by $1/z$ Da.

To obtain the number of signals in the LC/MS images, the number of proteins in the sample needs to be multiplied by the number of peptides (dozens per protein), the number of charge states of the peptide ions (around three), and the number of isotopes (around three).

Peptide signals have roughly the shape of a bivariate Gaussian distribution. However, departure from this ideal shape is common; there is often asymmetry in the elution profile, and a heavy tail of the distribution. A complete model of peptide signals is presented in Chapter 8 where it is used to study the set of peptide signals that can be detected by the proposed feature detection method.

Many peptides are not observable in LC/MS images, or with low-intensity. This is related to the ionization efficiency as explained in Chapter 2, and to spontaneous fragmentation during mass analysis. These low-intensity signals are of particular relevance for biomarker discovery, as it is expected that biomarkers belong to the low-intensity range.

Background noise Background noise in LC/MS images is composed of electronic noise and chemical noise. Electronic noise is created in the ion detector by stochastic variations of the current or other imperfections in the electrical appliances like the power supply [SMKM07].

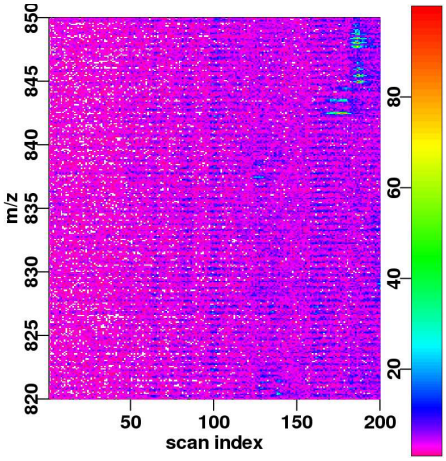

Figure 6.4: Background noise.

Chemical noise is the major source of background noise in LC/MS images. It is the limiting factor in the sensitivity of the mass spectrometer rather than the intrinsic capabilities of the instrument[SMKM07]. Chemical noise arises from ion fragments generated during ionization and mass analysis. As these are ordinary proteins, the behavior of background noise is similar to regular peptide signals [ARC+03]. For instance, both display a 1Da pseudo period that corresponds to isotope patterns.

In this thesis, background noise is studied in more detail in two different contexts. First, background noise adds a positive contribution to the recorded intensity. In Section 6.7, Section 6.8 and Chapter 7, we study background noise to improve the quantitative measurements acquired on an LC/MS platform. Second, background noise hides real peptide signals from feature detection and creates false positive identifications. We address this issue in Chapter 8.

Contaminants During the sample preparation procedure, some types of molecules are introduced unwillingly in the sample. These include keratins from skin and hair, plastic residues from the instrument tubing and other chemical compounds that are present in the lab atmosphere. Some contaminants form groups of small molecules that aggregate into larger complexes and fall into the mass range of the mass analyzer. [KSYW08] provide an extensive list of known contaminants in LC/MS images.

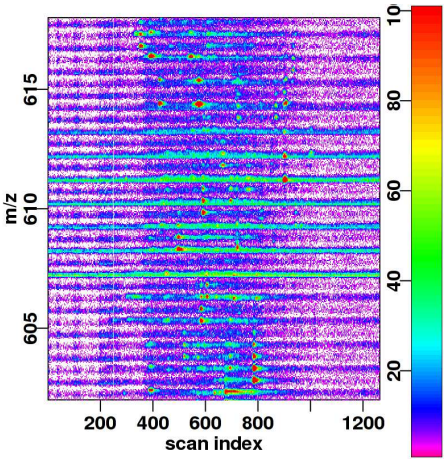

Figure 6.5: Contaminants.

Some of the contaminants are separated by liquid chromatography, and appear as regular molecular signals in the LC/MS image. Theoretically, it is possible to distinguish those from peptide signals based on the isotopic patterns because the distribution of intensity among the various isotopic forms reflects the chemical composition of the molecule. In practice, current mass analyzers do not have the necessary accuracy and resolution to perform that analysis on peptides. It would also be difficult to distinguish between an exotic isotopic pattern and overlapping peptide signals.

Other contaminants are easy to detect because they are not separated by liquid chromatography and are present in all mass spectra (at the same m/z ratio) as displayed in Figure 6.5

In this thesis, contaminants are not modelled explicitly. If they are separated by LC, then they are considered as standard molecular signals are reported in the feature detection results. If they appear as horizontal lines, they are included in the components of the noise model as chemical noise.

6.1.4 Measurement capacity of the LC/MS platform

LC separation For proteomics applications, most experimental protocols use a liquid chromatography column and a separation that lasts on the order of thirty minutes. Mass spectrometers acquire a mass spectrum roughly every second, which totals about 1,000 mass spectra in a LC/MS image in electrospray ionization (MS/MS spectra excluded). In MALDI ionization, the sample is traditionally deposited on plates with roughly a hundred spots. More fractions can be obtained by continuous deposition on the MALDI plate, see [WBF+02, ZNM04] for further details.

M/z separation Peptides obtained after tryptic digestion in sample preparation usually weigh less than 2,000 Da. On the other hand, a peptide fragment needs to be of sufficient length for identification of its parent protein. Five or six amino acid residues are enough, which corresponds to molecular weight higher than 400Da. Consequently, the mass spectrometer records intensity values for the signals in the m/z range $[400; 2,000]$. However, as mass analysis deals with m/z ratio instead of molecular weight, that range includes molecules up to 8,000 Da when considering ions with four charges.

Depending on the type of mass analyzer, the sampling rate on the m/z axis is variable; the number of sampling points in a mass spectrum ranges from 20,000 to one million intensity values. Sampling locations are reproducible between consecutive scans before calibration of the m/z values.

Mass analyzers have different sampling patterns. Ion traps and quadrupoles use uniformly spaced sampling. TOF instruments sample uniformly in the time domain, but after conversion to m/z values, the sampling rate is higher for lower m/z ratios. Fourier transform instruments sample uniformly in the time domain, and after transformation to the frequency domain, the sampling rate is uniform on the m/z axis.

Intensity measures As previously indicated, intensity values are positive real numbers. In `mzXML` files, these are encoded with 32bit precision floats, but underlying analog-to-digital converters only provide on the order of 16 to 20 bits of precision.

The observed range of intensity values is lower and spans approximately three orders of magnitude. There is a factor 10^3 between the maximum intensity and noise level. When computing the area under the curve, the mass spectrometer can report intensity values on approximately four orders of magnitude.

The detector is linear in a large range [3]. However, the molecules in the sample may compete in the ion source. Abundant molecules such as contaminants are known to mask the signals from low-concentration molecules, and lead to non-linear and saturation effects.

Conclusion A LC/MS image is a very large image. The data sets considered in the course of this thesis contain images with 80 million to one billion pixels, with several gigabytes in file size. LC/MS images contain large numbers of signals, as well as low-intensity signals that are relevant for biomarker discovery. In that context, data storage is a major issue; some labs can generate 10 GB every day. Computational time is also an issue, and we have consequently focused on efficient algorithms.

6.1.5 Compression and centroiding

There are few initiatives to compress LC/MS data, even for the open XML file formats that are not very efficient. Most of the community lacks computer skills and dedicated personnel, which would slow down the adoption of software dealing with compressed file formats like [MKW+06]. Standard algorithms (LZW, Bzip2, LZMA) can compress `mzXML` files reasonably well due to the nature of XML file formats. However, they do not take advantage of the specific structure of LC/MS data.

Centroiding is a major compression method for mass spectra. It essentially detects neighboring intensity values corresponding to the same peak, and aggregates them together. Only one intensity value per detected peak is recorded in the data file. However, centroiding is a default option and implemented in proprietary software, and the method implementations are not described in detail. Moreover, it disturbs the statistics of background noise and creates artifacts in the LC/MS image as shown on Figure 6.6.

[STL+06] actually uses the data voids created by centroiding to detect features in LC/MS images. In doing so, however, they entrust the complexity of detecting signals to the manufacturer software.

[3] except time-to-digital detectors as described in Section 3.4.

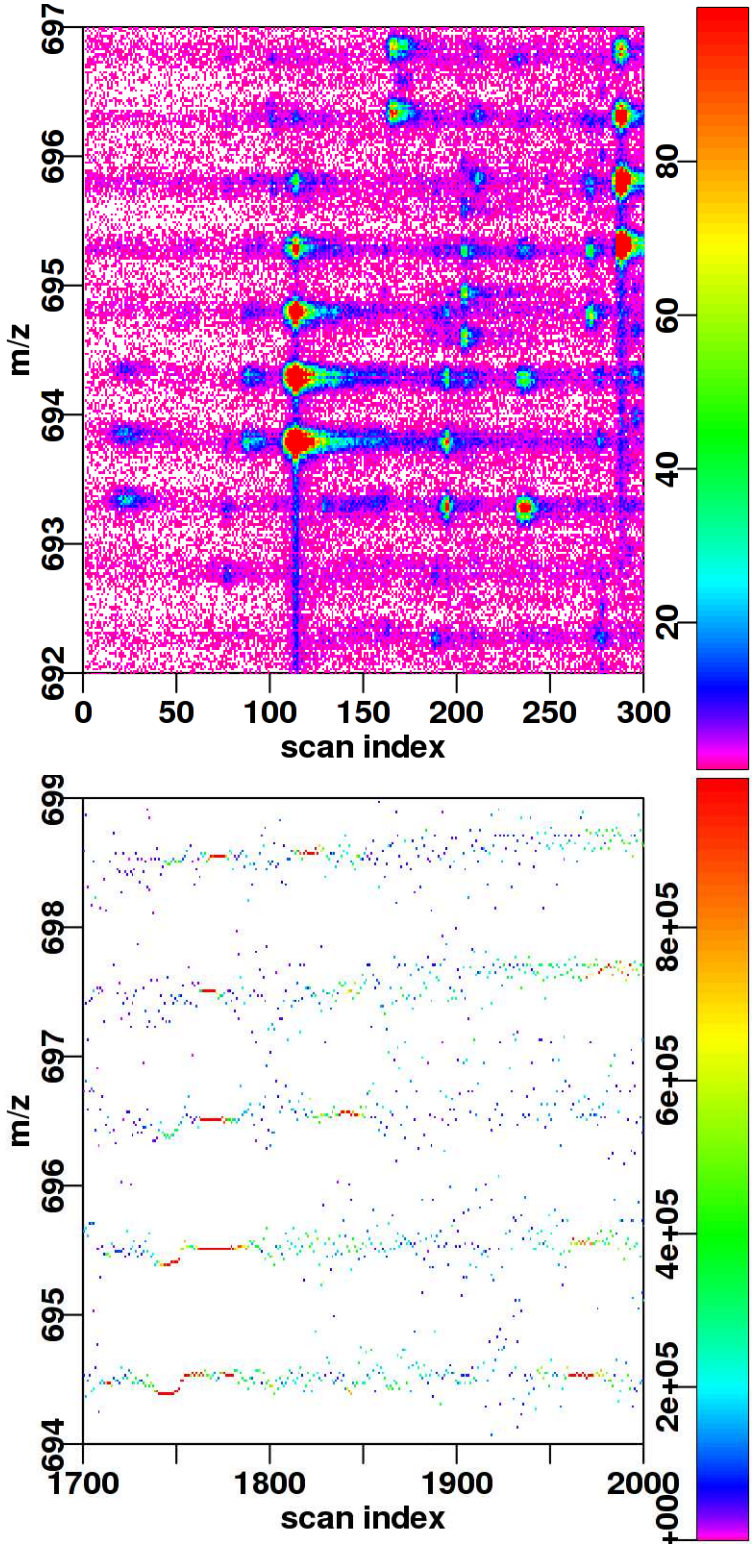

Figure 6.6: Comparison betwen LC/MS images from a TOF instrument (top) and an FT-ICR instrument with centroiding (bottom). Centroiding creates voids in the background noise around high intensity signals. It also changes the location of signals.

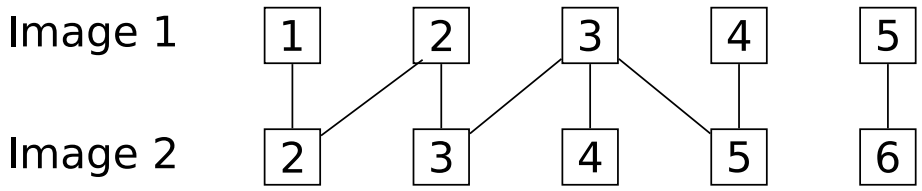

Figure 6.7: Multiple-to-one alignment of two series of mass spectra.

6.2 Retention time alignment (summary)

Liquid chromatography is notoriously known for its poor reproducibility, and the fact that retention times of peptide signals do not correspond in different LC/MS images as shown in Figure 6.8. Even so, Section 6.3 motivates the attempts to match retention times, and applications of these methods are described in Section 4.4 for protein identification and Section 5.2 for biomarker discovery.

Retention times can be matched across different experiments because the elution order of peptides is preserved. Using this hypothesis, many methods for image alignment have been imported from other scientific domains like signal and audio processing [SC78], image registration [ZF03] and 2D gels [VDY01]. Section 6.4 provides a discussion of experimental reasons and hypotheses for retention time alignment, and also provides a review of the current methods.

Sections 6.3 and 6.4 were extracted from my work published in [VLTTK+08]. Nevertheless, reading this paper is a daunting task for the uninitiated, and we introduce it with a detailed presentation of a typical alignment procedure in the remainder of this section.

6.2.1 Detailed example (ChAMS)

ChAMS (CHromatography Aligner using Mass Spectra) is a retention time alignment tool for LC/MS images developed by Amol Prakash under the supervision of Benno Schwikowski. I have had the pleasure to discuss at length with the author about the implementation in [PMW+06] and its application for quality control of experiments in [PPW+07].

When operating on LC/MS images, several types of information can be used for alignment: MS/MS identifications, MS signals, etc. ChAMS operates on the level of MS scans, and tries to find the correspondences between scans in two different images (binary alignment). By using the intensity measurements in the scans, ChAMS takes advantage of low intensity signals that are not processed: not detected as features, not selected for MS/MS identification, etc.

ChAMS produces a multiple-to-one alignment (see Figure 6.7), i.e. several MS spectra in a LC/MS image may correspond to the same spectrum in the other image. Other algorithms produce one-to-one alignments, allow mismatches or gaps. Some methods compute a transformation of the retention times without computing correspondences.

In order to find pairs of corresponding spectra, ChAMS computes a similarity score d between all possible pairs of mass spectra. This score is based on the Pearson correlation coefficient between the intensity values in the mass spectra. To address m/z calibration problems, correlation is computed with a tolerance ε in the m/z value.

More precisely, let $f(m)$ and $g(m)$ denote the intensity as a function of m/z ratio in two mass spectra. [PMW$^+$06] defines the intermediate score

$$d(f,g) = \sum_{|m_1 - m_2| < \varepsilon} f(m_1) g(m_2)$$

From this score, the following normalized score is computed

$$s(f,g) = \frac{d(f,g) - \mathbb{E}[d(f,g)]}{\sqrt{d(f,f)\, d(g,g)}}$$

where $\mathbb{E}[d(f,g)]$ is the expectation of $d(f,g)$ under random permutation of the intensity values inside the mass spectrum.

The similarity scores for all pairs of mass spectra are assembled in an alignment matrix as shown on Figure 6.11 on page 104. In that matrix, high similarity scores indicate pairs of spectra that belong together. Retention time alignment consists in finding a path in the matrix that goes through the matrix cells containing the highest scores. This optimization can be computed efficiently using dynamic programming algorithms.

As detailed in Section 6.4, most methods for retention time alignment follow a similar pattern. They are characterized by the type of input data used for alignment (MS spectra for ChAMS), the choice of possible alignment transformations (multiple-to-one alignment), the scoring function used (s), and an optimization procedure to find the best transformation (dynamic programming).

6.3 Retention time alignment (Introduction)

LC-MS-based proteomics is an analytical approach for the analysis of complex peptide or protein mixtures [PG01]. Compared to gel-based proteomic approaches, LC-MS allows online separation, high sampling rates, and easy automation, together with a high level of sensitivity [BS05].

A typical LC-MS-based proteomics experiment involves the following series of steps. First, proteins are digested into smaller peptide fragments using an enzymatic reaction. Secondly, the resulting peptides are separated using liquid chromatography; separation is usually obtained by hydrophobicity (reverse-phase, RV), but other characteristics like charge (strong cation exchange, SCX) are used also. HPLC columns are particularly effective devices for this task; they allow separation of thousands of peptides. Each resulting fraction corresponds to a mixture of peptides having similar separation characteristics. According to these characteristics, fractions elute from the column at different time points.

In the third step, the eluting sample is injected into a mass spectrometer, where the constituent molecular masses are recorded. Typical separation times range around 60-100 min, and recording a mass spectrum approximately every few seconds yields a sequence of several thousands of mass spectra.

The resulting data can be represented as a 2-D image (Fig. 6.8), where the horizontal axis represents the elution time (retention time; the time at which an HPLC fraction was acquired by MS), the vertical axis represents the m/z in the corresponding spectrum, and color represents intensity in the mass spectrum. For a comprehensive review of these and other aspects of MS-based proteomics, we refer the reader to the overview by Aebersold and Mann [AM03] and two books [SKG97, McM07].

The analysis of an LC-MS image requires a number of preprocessing steps, such as calibration and quantization of m/z values, filtering, noise reduction and background subtraction, amplitude normalization, peak detection and peak quantification. Preprocessing has been studied extensively, and we refer the reader to two reviews [LE05, RJR$^+$04].

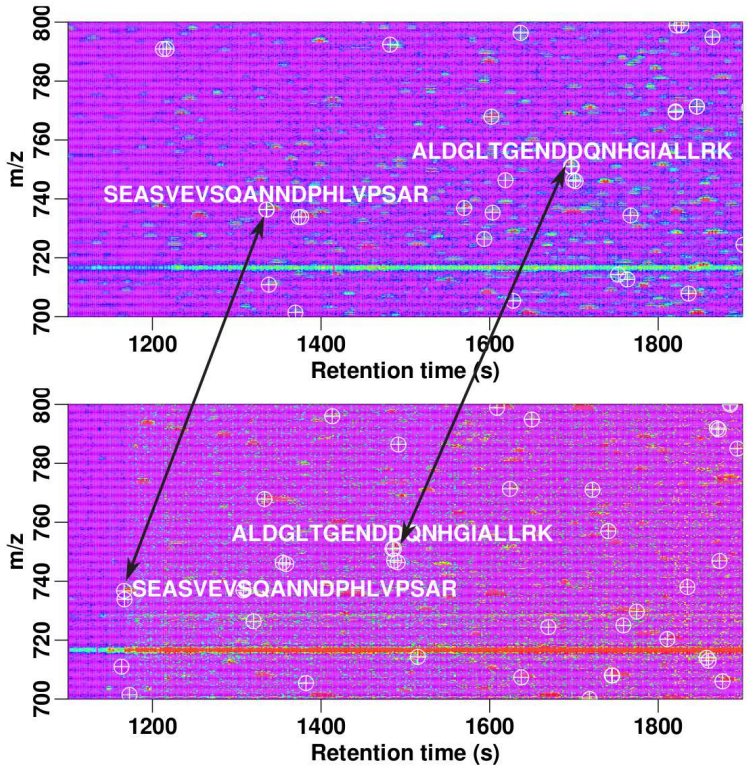

Figure 6.8: Alignment of two LC-MS images using MS/MS identifications. Images were acquired for two similar experiments on an LTQ-FT mass spectrometer by Guina et al. [GRB+07]. High confidence MS/MS identifications with protein p-value 0.01, peptide score > 20 and rank 1 were obtained using Mascot. Identifications apparent in both images are marked with cross-hairs. Aligning these identifications exposes a significant retention time shift of the lower image compared to the upper image; example peptide matchings are shown by the two arrows.

After preprocessing, the image is subjected to peptide and protein identification algorithms, which use protein sequence databases or previously acquired identifications to compute a list of proteins contained in the sample [AM03]. Depending on the experimental question, quantification algorithms for estimating the abundance of identified proteins can additionally be applied [OM05, MCK+05].

While the above preprocessing steps only deal with one LC-MS image at a time, many questions in biological and biomedical research involve the comparison of proteomes of different samples, and thus, multiple LC-MS images. For example, in biomarker discovery, one is interested in finding specific peptides that allow distinguishing between healthy and disease states. Comparing multiple LC-MS images is difficult; even identifying a consistent set of peptides in repeat experiments on the same sample is challenging [CM05]. In particular, comparing LC-MS images solely on the basis of lists of identified peptides provides only a qualitative and coarse basis for comparing the underlying biological states and is thus insufficient for most applications.

Alternatively, one can attempt to compare the data-rich LC-MS images directly. A major technical difficulty in the direct comparison is the notorious irreproducibility of the LC separation step [SKG97, McM07]. Even in technical repeats of the same experiment, any given peptide typically does not elute in the same fraction(s) [VY04] (Fig. 6.9). This irreproducibility shows up in LC-MS

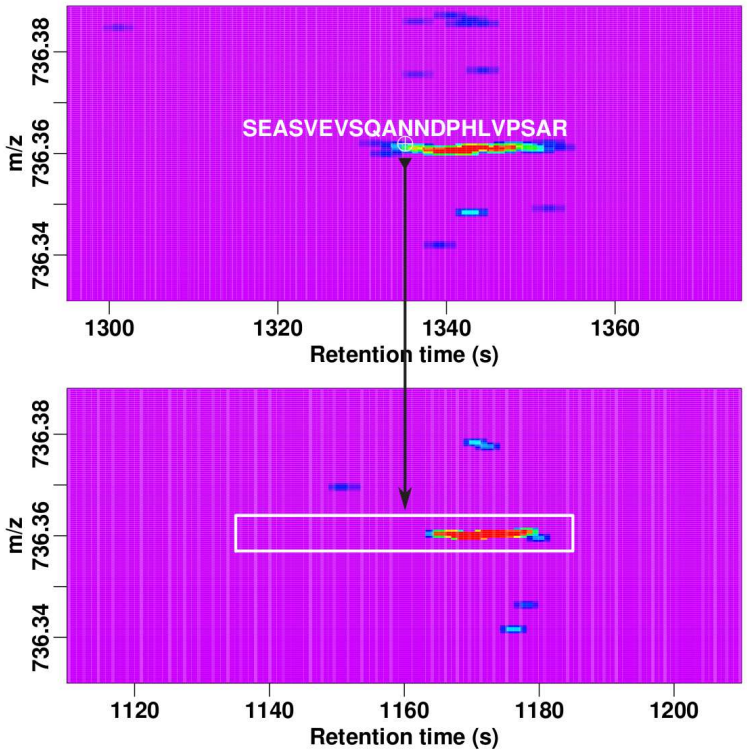

Figure 6.9: Example of peptide identity propagation. The figures show a small window of the data presented in Fig. 6.8. After alignment of the two images, the MS/MS identification in the upper image (cross-hair) can be propagated to corresponding MS features in the other image. A white rectangle denotes the valid range of putative corresponding MS features.

images as distortions of the elution time axis, and necessitates alignment, a corrective transformation of the elution time axis between two or more images. Only after this transformation can individual features in different experiments be compared directly.

Several biochemical differential labeling approaches have been developed to circumvent the problem of LC irreproducibility and to allow comparison of different samples in one single LC-MS experiment [GRG$^+$99, Man06, OBK$^+$02, JR04, JR05, LSG03]. The main idea is to tag peptides or proteins of different samples by chemically attaching labeled molecules. For each sample, a label with a distinct mass is used, each resulting in a known modification of any given peptide's molecular mass. After mixing the samples and recording a single LC-MS image of the mixture, corresponding peptides from two samples can be identified as peak pairs in the mass spectrum with a distinct mass difference. Since only one LC-MS image now contains the information for all samples, no alignment between experiments is necessary.

However, differential labeling has several limitations. First, the requirement that all samples need to be available in the same location at the same time restricts the applicability of this approach. Moreover, due to the restricted number of possible labeling reagents, only a limited number of comparisons are possible; furthermore, reagents may also constitute a significant cost factor. Finally, sensitivity is typically reduced by the required biochemical preprocessing steps and the limited peak capacity of the mass spectrometer.

The growing number of publicly available LC-MS datasets increases the interest in comparing datasets generated in different laboratories. Label-free LC-MS analysis provides an alternative to differential labeling that avoids many of the abovementioned issues [OMAAW+05]. Here, each sample is subjected independently to LC-MS measurement; the multiple LC-MS images are compared to each other and analyzed for their similarities and differences. In particular, label-free analysis does not require a mixing of samples and thus enables comparison of data sets generated in different laboratories or at different times. Recent progress in the accuracy of mass measurements enabled the development of accurate and efficient alignment methods to correct for the differences in elution times. These methods are a key technique for labelfree analysis of multiple LC-MS images.

The ability to make label-free comparisons gives rise to several new applications. By combining the information of several LC-MS experiments, the coverage of identified and quantified peptides and proteins in a given sample can be increased. Although most peptides and proteins in complex mixtures remain unidentified in single LC-MS/MS experiments [LYK+05], their identity can frequently be inferred by associating an already identified feature in one image to a corresponding unidentified feature in another image (Fig. 6.9). Biomarker discovery is another important application area of comparative LC-MS-based proteomics. After preprocessing and detection of features, such as chromatogram peaks or other image patterns, biomarker candidates are identified as those features that can reliably discriminate between the two sets of experiments. Clearly, alignment of LC-MS images is a key component in this process. Usually, alignment also provides a measure of similarity of LC-MS images; this measure can be used to establish a quality control for LC-MS procedures and allows the characterization of technical and biological variation in LC-MS images across experiments.

Since variations in the LC separation step have proved to be hard to avoid in practice and due to the increased interest in label-free methods, research on alignment approaches is an active area in computational MS. This review aims at providing an overview on methods for elution time alignment and their main applications. The manuscript is organized as follows: we motivate and define the problem of elution time alignment, survey and classify the main approaches. Additionally, we give some guidelines to facilitate the practical choice of an alignment method, and give some pointers to freely accessible LC-MS image comparison software (Section 6.4). We then discuss alignment approaches in two major application contexts: protein identification (Section 4.4) and biomarker discovery (Section 5.2). We would like to point out that we tried to keep Section 6.4, Section 4.4 and Section 5.2 as mutually independent as possible and thus, minor repetitions in discussing methods seemed unavoidable. Finally, we offer our personal view of the key future trends and developments in the field (Section 6.5).

6.4 Computational Approaches to Elution Time Alignment

In this section, we review technical approaches to elution time alignment. Elution time alignment approaches have been the topic of two previous reviews by Tomasi et al. [TvdBA04] and van Nederkassel et al. [vNXL+06]. Since their publication, a large number of new approaches have been published in the literature; this section is a significant extension and an update of these reviews. It is organized as follows: we start by discussing the causes of experimental elution time variations in LC columns that motivate the need for computational alignment techniques. We then compare LC-MS alignment to related computational problems, namely image registration and 2-D gel analysis. In Section 6.4.1, we review and categorize alignment approaches based on their type of input. For a deeper technical understanding, alignment approaches are characterized by five different characteristics that characterize strengths and limitations of each approach. Section 6.4.2 discusses these characteristics individually, as they can be interchanged between approaches to create new combinations and to extend existing approaches. Section 6.4.3 discusses

how elution time alignment approaches have been compared and validated, a difficult and important issue for most large-scale experimental platforms. In Section 6.4.4, we summarize our advice on the problem of choosing an alignment method in current applications.

Elution Time Variation

A considerable amount of variation in the LC separation originates from the HPLC column itself. Major causes include aging, packing, and contamination of the column [McM07, TvdBA04, vNXL+06, Eil04, PKF+03, WW05]. It is worth noting that these problems are column-specific and occur even in highly controlled experimental conditions. As an example, a biomarker study that involves ten controls and ten cases with three replicates uses 10% of the typical column life span (60 runs out of a few hundred). Additional variability is expected due to other experimental factors that are difficult to control, such as temperature, gradient shape, and mixing physics [vNXL+06, WW05, JMP+06, NMA+05, PMW+06]. The chromatography may also change when a column is replaced. Finally, instrument malfunction is another source of problems; specific examples include improper pump function [ACvG+06], dead volumes, flow rate [NMA+05, PMW+06], leaks, and drift [WW05].

Reproducibility of other chromatography setups, such as gas chromatography (GC), suffers from similar problems; among them are oven temperature control, the GC column, etc. [KTZ+06]. Some of the algorithms described here for LC-MS data were originally developed for aligning GC-MS data [TvdBA04, vNXL+06, KTZ+06, Eil03], but are applicable to LC-MS data without modification.

Elution Time Alignment

As elution time variation is hard to eliminate, the measured elution time cannot be used directly to compare different experiments. Rather, elution time axes have to be mapped onto each other first, a procedure commonly referred to as alignment.

We use the term "alignment" in this more relaxed sense compared to, for example, classical sequence analysis. Mathematically, an alignment in the strict sense is a mapping of two sequences that preserves the order of elements in both sequences. As an example, alignment algorithms for DNA sequences only allow sequence insertions, deletions and matches/mismatches between sequence elements, but no sequence inversions, which can also be observed in DNA sequences. Inversions do not preserve the order of the sequence, and therefore require a more relaxed formalism.

Similarly, although the exact shape of a required corrective transformation of the elution time axes in LC-MS experiments is generally unknown and experiment-specific, it appears likely that the peptide elution order is preserved for the same type of column [KTZ+06]. Nevertheless, local inversions of elution order are observed, but seem to occur rarely [KSS+07]. Additional peptide-specific elution time corrections are therefore sometimes applied after a global transformation of the elution time axis, and considered part of the alignment. Corresponding to the more relaxed use in the literature, we use alignment in this review for a general map between different (m/z, elution time) windows.

It is worth noting that, although elution time alignment is distinct from m/z calibration, the computational approaches for both problems are somewhat similar, and many papers on m/z calibration actually use the term alignment [THNC02, Jef05, RY06]. Mass measurement quality is largely defined by the physical capabilities of the mass spectrometer, although sample preparation and in particular flow rate (sample concentration in the eluent flux) do affect subsequent analysis, because of competitive sample ionization and space-charge effects in Ion Trap (IT) and FTICR mass analyzers. With the advent of high-accuracy mass spectrometers, mass accuracy

and calibration have become less of a problem and we will neglect these issues throughout the manuscript.

Relation to Image Registration and 2-D Gel Analysis.

As LC-MS experiments are naturally displayed and modeled as 2-D images, approaches for their alignment can be compared to image registration and 2-D gel alignment. In all cases, an alignment algorithm seeks for a transformation that "best" maps one image onto the other. Different assumptions about the causes for the differences between the images can be described by different sets of transformations, wich in turn lead to different alignment algorithms.

In image registration, images are assumed to contain identical objects that have been recorded from different camera perspectives; mathematically, they reside in different coordinate systems. Two such coordinate systems can be transformed into each other using a linear transformation in three dimensions; however, this (unknown) transformation may result in corresponding non-linear transformations between the 2-D images, due to projection on the focal plane of the camera. To compute these transformations, the images are typically separated into the relevant foreground objects, and the background. Boundaries are estimated by segmentation of the image into large areas of constant color, light intensity, or identical texture, and foreground objects are identified from these boundaries. Once the foreground objects are found, an optimal mapping between corresponding objects in the two images is computed. Image registration has important applications in medicine, cartography, and computer vision, among others [ZF03].

Although any image registration tool can also be used for elution time alignment, it is likely to perform poorly, mainly because a clear separation into fore- and background is hard to achieve for LC-MS images. In particular, LC-MS images are composed of spots on a noisy background. Segmenting an image by finding object boundaries would result in localized and dense clouds of points. Classical image registration methods would interpret these clouds as background or noise artifacts.

In 2-D gel analysis, images are obtained from experiments using 2-D PAGE. Features are usually larger clouds of dense pixels, typically corresponding to stained (complexes of) proteins. A 2-D gel alignment method aims to map corresponding features between images [VDY01, DDY03]. Differences are explained by elastic deformations of the medium that occur in both directions [ASNN05]; they can be modeled with strong regularity constraints on the feasible transformations. Image registration methods are much more suitable for 2-D gel alignment than for LC-MS alignment, as prominent features are more easily identified as foreground due to their size. Robust image registration is thus an essential prerequisite for subsequent quantification procedures [VDY01].

6.4.1 Alignment Approaches

All alignment approaches fall into one of two very broad categories: They are based either on profile data, where the LC-MS images are taken as recorded, or on feature data, where the data have been processed to identify important signals and distinguish these signals from noise.

Profile-based approaches use the full, unprocessed data obtained in the LC-MS experiment, and usually attempt to find an alignment, such that the overall difference between intensities of two LC-MS images after alignment is minimized. While small, unavoidable, differences due to random intensity variation between experiments are expected, large differences are assumed to result from misalignments of unrelated parts of the experiment.

Feature-based approaches, on the other hand, explicitly try to distinguish between relevant signals (features) from peptides and irrelevant noise parts in the data in an initial feature detection step, and rely only on these features for the subsequent alignment. Feature detection typically excludes a large amount of noise and thereby reduces the data considerably. Clearly, the performance of the overall alignment process strongly depends on the performance of the feature detection step and we refer the reader to [LGR+06, KO05] for references in feature detection methodologies. In contrast to profile-based approaches, feature-based alignment methods typically take mass differences of aligned features into account; only few of them also use signal intensity.

While profile-based alignment methods are able to perform well on data with low m/z resolution (e.g. LTQ data), feature-based methods usually require accurate mass measurement for successful feature detection.

6.4.1.1 Profile-based Methods

The most basic profile-based methods compare only the difference in the TIC (total ion chromatograms), and thus make no or little use of the full LC-MS image. We first describe these methods, which are thus only suited for low-complexity data, before the methods that exploit the higher resolution of LC-MS images.

Aligning TIC

Technically, a TIC profile is a sequence that represents the total ion count in each fraction of the LC separation. Most methods for aligning TIC profiles are inspired by the original work of Sakoe and Chiba [SC78] for speech processing and the ensuing developments; these methods are technically similar to biological sequence alignment [Gus97, SK83]. The two sequences of ion counts (TIC profiles) (m_i) and (n_j) of two LCMS images are matched using a pairwise distance function $d(.,.)$. This distance function assigns a score to each possible pair (m_i, n_j) of ion counts using their intensity difference, e.g. $d(m_i, n_j) = (m_i - n_j)^2$.

An alignment corresponds to an order-preserving path, a sequence of pairs (m_i, n_j) where each m_i and n_j occurs in the same order as in the TICs (m_i) and (n_j). The score s of an alignment is computed from the pairwise distances by summation over constituent pairs (m_i, n_j): $s = \sum d(m_i, n_j)$. The choice of a suitable distance function depends on the application context, and represents a major difficulty in all dynamic programming-based approaches. Once the distance function is chosen, an optimal alignment can be found efficiently (in time quadratic in the length of the TIC) using the dynamic programming technique, the corresponding alignment approaches for time series is called dynamic time warping (DTW).

The alignment score function can also be designed to introduce additional assumptions. As in sequence alignment, for total ion chromatograms, one can use a weighted sum of distances, penalize gaps, or restrict path jumps. An additional way to constrain the alignment path, called correlation optimized warping (COW), is proposed by Nielsen et al. [NAH+02]. Instead of evaluating all possible pairs (m_i, n_j), the authors first divide the chromatograms into non-overlapping segments. The segment boundaries are then adjusted by maximizing Pearson's correlation coefficient between pairs of segments. When segments of different length are matched in the alignment, one of the two chromatograms is interpolated to adjust the elution time axes before computing the correlation coefficient.

Bylund et al. [BDMM02] propose several modifications to the COW approach. They allow user-guided segmentation of the elution time axis, which allows the use of key features in the data to guide the alignment. They also include flexible matching of start and end points, and use

covariance instead of correlation as matching criterion in order to put more emphasis on large peaks.

Eilers et al. [Eil04] propose parametric time warping (PTW), an approach based on a parametric model of the alignment transformation. This model consists of polynomials of order less than some constant k; in their study on GC-MS data, quadratic polynomials ($k = 2$) were sufficient. The model coefficients can be efficiently computed by solving a least squares problem, which permits speed improvements over DTW.

Van Nederkassel et al. [vNXL+06] propose an extension of PTW called semi-parametric time warping (STW). Here, Bsplines of varying number and extent are used instead of polynomials of order k. The optimal coefficients are again computed by linear algebra.

Listgarten et al. [LNRE05] propose a generative model of a TIC based on continuous hidden Markov models (HMM). HMMs use an implicit reference model (a latent trace) that can be interpreted as a consensus TIC profile. The a priori likelihoods of specific distortions of this consensus TIC profile are explicit model parameters.

Higher-resolution profile alignment

In contrast to TIC-based methods, several recent approaches exploit the information available in full mass spectra to deal with mixtures that are more complex by separating the m/z axis into several mass bins.

In an extension of their previous HMM work [LNRE05], Listgarten et al. [LNR+07] use four mass bins for alignment. Wang et al. [WZL+03] use 200 m/z locations in each scan to compute the similarity between m_i and n_j in their DTW algorithm.

Block matching algorithms are tools from image analysis of video sequences used to estimate the motion between successive images. In [PBG+07], Paulus et al. propose to apply the fast block matching algorithm in [Gha90] to align LC/MS images.

Pursuing the idea to leverage as much information as possible, Prakash et al. [PMW+06] use spectra acquired without any prior data reduction in a DTW approach. Pairwise distance between MS scans is computed by multiplying the intensities of sufficiently close peaks in the two images and summing over the products. They then employ a compensation term for random matches and normalize the score.

To make the method more robust, neighboring scans are also considered in the score function. Prince and Marcotte [PM06] also use unprocessed spectra in a DTW approach to compute an initial alignment. This alignment is then analyzed to extract bijective anchors, which are interpolated into a bijective, smooth warp using a method outlined in Fritsch et al. [FC80]. The anchor selection favors areas of greater similarity between experiments. In their application study, four similarity score functions are compared: Euclidean distance, dot product, covariance, and correlation; the correlation yields the best results. The authors argue that, since bijective pairwise alignment is symmetric, it is favorable as a basis for multiple alignments.

6.4.1.2 Feature-based Methods

In data analysis, a feature is commonly a subset of the data with specific characteristics; in MS, features are usually peak models that summarize the distinctive characteristics of peaks, such as height, S/N, or shape in the LC-MS image. In this Section, we discuss approaches to align lists of features that have been extracted from the LC-MS images in a preceding feature detection step.

A single peptide typically generates several peaks that correspond to its isotopes [WTF+07] or different charge states. Since a peptide elutes over a certain time interval, its peaks typically appear in several successive scans. Certain feature detection methods will group all of these peaks into only one feature [LYK+05, JML+06]; feature overlap or p-values are used to assess the confidence in each grouping. We proceed to describe feature-based alignment methods by increasing level of resolution at which the experimental data is exploited for alignment.

Aligning chromatogram peaks

The profile-based methods, discussed in Section 6.4.1.1, treat each section of chromatogram equally, or biased according to weights that are attributed to individual sections. In the latter case, the alignment of certain more informative regions may therefore be obvious, whereas the alignment of peak-free regions may be uncertain.

Krebs et al. [KTZ+06] define "landmarks" to be LC peaks above a threshold on the TIC profile, and match those landmarks between two experiments if the correlation score of the corresponding mass spectra exceeds 0.99. The alignment is then computed by interpolating gaps between landmark peaks using cubic splines.

Walczak et al. [WW05] focus on finding correspondences between the LC-peaks in two experiments. Similarity between the peaks is computed using a Gaussian error model and summarized into a correspondence matrix. Fuzzy matching of peaks is performed by alternating between estimation of the correspondence matrix and shifting the peaks to improve matching. After convergence, the alignment is finalized with piecewise linear interpolation.

Aligning image features

Katajamaa et al. [KO05] build a "master peak list" of detected features in the LC-MS images. Each feature is either aligned to an existing entry in the master peak list if it is close enough to an existing entry, or appended to this master peak list. The distance between features is determined as a weighted sum of absolute deviations in m/z and elution time.

Radulovic et al. [RJR+04] note that assigning a single elution time value to a feature is difficult, since peptides generally elute over several fractions. Consequently, they assign both an m/z and an elution time interval to each peptide elution profile; each profile thus corresponds to a rectangle in the LC-MS image, called a "pamphlet". Accelerated Random Search [KRR+03] is used to compute a piecewise linear warping that optimizes rectangle overlap of pamphlets between experiments. Finally, a local relaxation is applied to improve the alignment.

Jaitly et al. [JMP+06] model variability in measured elution time by a Gaussian distribution with small variance, and allow a nonlinear warping (or trend) on a larger scale. Their alignment algorithm is a DTW approach with a distance function based on matches between peak lists. After (m/z, elution time) features are picked, the dataset is divided into N segments and dynamic programming is used to find matches between these segments by minimizing the distance between possible common features (weighted using the Gaussian model). Afterwards, individual features are grouped.

Kirchner et al. [KSS+07] use robust point matching [CR03] to find matches between peak lists. Matches are estimated using a fuzzy correspondence matrix similar to Walczak et al. [WW05], and outliers are treated explicitly. Given the correspondence matrix, the optimal warping function is found by smooth monotone regression using the Newton-Raphson algorithm. A deterministic annealing scheme is introduced to give more flexibility to the choice of the weights.

Wang et al. [WTF+07] use "peptide elements" that represent the expected signals generated by a single peptide. Each peptide element consists of a theoretical distribution of one unit of intensity over a certain (m/z, elution time) rectangle. A peptide element library (a set of possible peptide elements) is built from the given set of LC-MS images before alignment. Given this library, a robust regression scheme is used to find the peptides present in each scan, together with their quantity. For alignment, a group of common peptides is identified as a subset of the library that minimizes a statistical "loss function".

Silva et al. [SDD+05] and Mueller et al. [MRS+07] initially group features from two LC-MS images if they fall within a tolerance window of wide m/z (0.05 Da) and elution time (5 min). The assumed trend in elution time shifts between grouped features is estimated using LOWESS [Cle79] and related peptide features are merged. During the following iterations of this step, elution time tolerance is gradually reduced to 30 s. To align multiple LC-MS images with this

technique, Mueller et al. propose to build a consensus "MasterMap" using hierarchical clustering. LC-MS images are merged hierarchically, based on their overlap (Spearman correlation) and reproducibility of feature intensity.

Lange et al. [LGST+07] align images using affine linear transformations on the whole LC-MS image, thus including alignment of the m/z axis. As this class of transformations is defined by only two parameters, these can be found by tabulating likely parameter combinations from small subsets of data; parameters that occur most frequently are assumed to be correct. In a second stage, features known to correspond to each other (e.g. pairs of MS/MS identifications) are used to refine the parameters. Similar to Mueller et al. [MRS+07], the method is extended to allow alignment of multiple images by progressively aligning images to a reference.

Note that for some applications, like biomarker discovery, it is usually sufficient to only compute a coarse elution time correction to enable correct clustering of related features [KO05, JML+06]. However, in many other applications, like identification transfer or comparative protein profile studies, accurately correcting the images is crucial [JMP+06, KSS+07, SDD+05, MRS+07].

Aligning significant features

Some feature-based methods place an emphasis on "significant" features in LC-MS images, which are features that are more likely to correspond to actual peptides than other features, or even have additional information associated with them. Significant features are expected to be more reliable for aligning; examples are high-intensity peaks, peaks that are better resolved, or peaks with a corresponding MS/MS identification.

Zhang et al. [ZAA+05] define significant features as local maxima in a (m/z, elution time) window that are consistently present in all data sets. From multiple LC-MS images, a reference is chosen as the image that is closest to a consensus of the significant peaks in median elution time. Robust linear regression on the significant peaks is used to find a straight-line alignment, which is subsequently improved by small-scale deviations.

In a protein profiling study, America et al. [ACvG+06] use a commercial tool called MetAlign. This software estimates the baseline and noise characteristics in mass bins before detecting peaks. The algorithm starts from landmark peaks defined as features with a high S/N present in all datasets, and iteratively includes other features in the order of decreasing intensity to optimize a non-linear alignment path.

In SpecArray, Li et al. [LYK+05] compute a retention-time calibration curve (RTCC) between two LC-MS images based on pairing features previously detected by the included PEPLIST program. Detected features are loosely matched, and the RTCC is iteratively refined by minimizing the root mean square distance between paired features and removing aberrant pairs based on a Gaussian model. Features are successively removed from a super-set, whereas, in most other iterative approaches of this type, features are iteratively added to a consensus set.

The PEPPeR system by Jaffe et al. [JML+06] performs peak clustering and peptide identification after a coarse correction of elution time. High-confidence MS/MS identifications are matched to MS/MS identifications from a reference run, and the elution times are corrected with a quadratic polynomial. Aberrant corrections due to incorrect MS/MS identifications necessitate an additional outlier rejection scheme.

Fischer et al. [FGR+06] use all MS/MS identifications to compute a high-confidence initial alignment using ridge regression. This robust regression technique yields an optimal polynomial elution-time transformation by solving a reweighted least-squares problem. This polynomial is then iteratively refined using features that correlate best with the current alignment.

Instead of using all MS/MS identifications, Petritis et al. [PKF+03] initially use only six peptides identified by MS/MS across all experiments. Using a genetic algorithm, they find linear transformations that concurrently minimize the variance of the normalized elution times (NET)

across all experiments. Petritis et al. apply their method to align 687 LC-MS images from different bacterial species.

6.4.2 Five characteristics of Alignment Methods

Any alignment approach can be analyzed in terms of five different characteristics that largely determine its principal strengths and limitations. Together with the classification into profile- and feature-based alignment, these characteristics are as follows:

1. Type of input data used
2. Feasible transformations of the elution time axis
3. Scoring of alignments
4. Algorithms for computing the optimal alignment,
5. Choice of a reference experiment (in progressive multiple alignment).

Each characteristic is largely independent of the other, and interchangeable within each class. They may thus be combined to better suit the needs of a particular application scenario. In particular, the type of input data used makes a clear difference for many applications. Certain publications explicitly discuss available choices with regard to specific characteristics [PM06]. We proceed to discuss each characteristic in turn.

6.4.2.1 Type of input data

As outlined in Section 6.4.1, three types of data can be used for alignment:

- general profiles and TIC,
- LC-MS image features,
- and MS/MS identifications.

TIC provides the coarsest type of information; it is usually sufficient in GC-MS and LC-MS of low-complexity samples. On more complex samples, TIC information is less easily interpreted because peaks can overlap, and unrelated peaks can elute at a similar elution time [KTZ+06]. Multiple TIC approaches and full-profile-based approaches can exploit more detail and are particularly robust, but these approaches also need to deal with a higher level of noise, as low-intensity peaks in LC-MS images are often not recognized as features, or selected for MS/MS measurement.

The use of LC-MS image features to compute the alignment can be viewed as a compromise between TIC and MS/ MS-based alignment. On one hand, LC-MS image features are much more numerous than MS/MS identifications, but, on the other hand, they provide more precise information than the TIC. For a reliable and robust alignment, however, accurate mass measurement and high-resolution LC separation seem crucial for this approach.

MS/MS identifications provide a reliable way to assign corresponding peaks. However, they are not always available in the data, and, often, only a fraction of all MS/MS spectra will lead to high-confidence identifications. Mueller et al. [MRS+07] also report that 95% of MS/MS identifications belong to an LC-MS image feature, but that only 10% of such features are analyzed using MS/MS. Moreover, usually only a subset of MS/MS identifications is common between two experiments, and can be used for assigning feature correspondences.

However, the uses of LC-MS image features and MS/MS identifications for alignment are not mutually exclusive: Fischer et al. [FGR+06] show that their alignment can be improved by refining the alignment initially obtained from reliable MS/MS identifications with LC-MS image features, which provide more data for the alignment process. Lange et al. [LGST+07] take the opposite approach and prioritize features before refining using MS/MS identifications.

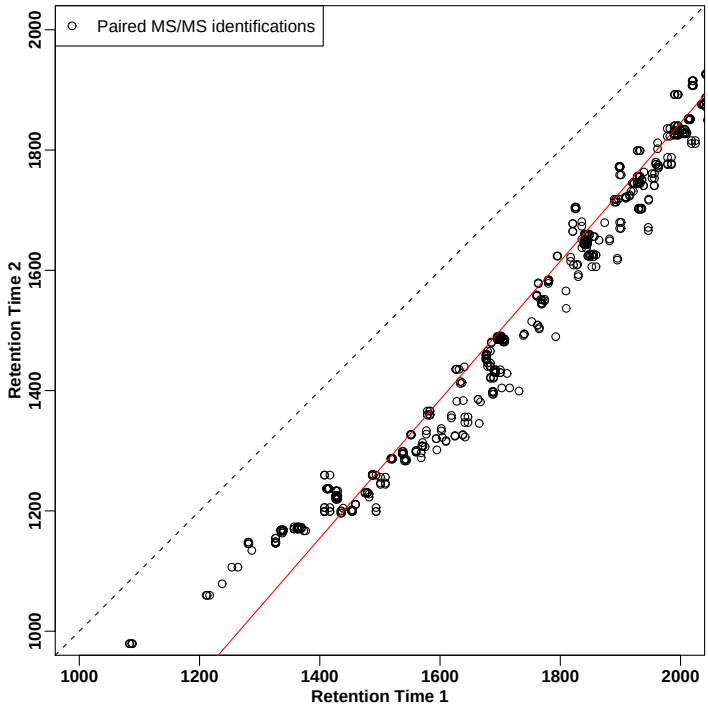

Figure 6.10: Peptides do not elute at the same time in different experiments. Each circle in the figure represents a pair of high-confident common peptide identifications in the two experiments. For a perfectly reproduced LC run, circles would be expected to lie on the indicated dashed line. However, linear regression (red line) suggests that there is a 6-min-offset between the elution times, and that elution speed is about 15% higher in the second experiment; elution time variability is around 2 min. The data also indicates that the elution time distortion is not linear; this behavior might thus be better captured using non-linear regression techniques.

6.4.2.2 Feasible Transformations of the Elution Time Axis

Another important characteristic that varies significantly between approaches is the set of considered corrective transformations of the elution time axis. The set of possible transformations correspond to the anticipated degree of complexity of the distortions in the HPLC separation.

Highly reproducible chromatographic platforms need only corrections for constant or linear time shifts (cf. Fig. 6.10), corresponding to static differences in acquisition start and flow speed. However, when comparing different elution gradients, one may expect more complex distortions [JMP+06]. Polynomial constraints allow more flexible alignments [Eil04, FGR+06], although van Nederkassel et al. [vNXL+06] suggest that quadratic polynomials are limited in flexibility, and not sufficient to model complex distortions.

It is generally desirable to keep the set of feasible transformations as small as possible to avoid overfitting the alignment to artifacts in the data [TvdBA04, PM06]. Van Nederkassel et al. [vNXL+06] report that the number of B-splines in the STW warp path has to be lowered to remove artifacts in the alignment. Similarly, Tomasi et al. [TvdBA04] note that COW gives better alignment results than DTW because of stricter constraints, and results become comparable only with the use of very rigid constraints in DTW.

When using landmarks or significant features, the space of possible transformations corresponds to the possible matchings between the high-confidence feature lists. To reduce the rate of false assignments, distance constraints between matches are usually applied [JMP+06, ACvG+06, KO05, MRS+07]. The resulting transformation is then computed using, for example, linear [RJR+04, WW05, BDMM02], or cubic spline interpolation [KTZ+06].

Many constraints reflect the assumption that the elution speed is roughly comparable between experiments. The constraints are implemented using gap penalties in dynamic programming alignments, constraints on dilatation/compression [TvdBA04, JMP+06], slope [PM06], local continuity [TvdBA04], or introduction of slack parameters [NAH+02]. Instead of local constraints, Kirchner et al. [KSS+07] bound the overall acceleration implied by the transformation.

6.4.2.3 Score functions for evaluation of possible alignments

Clearly, the available input data will determine the general type of score function that can be used. With TIC information, most approaches use a variant of Pearson's correlation coefficient. For more general profile data, Prince and Marcotte [PM06] present a systematic comparison between different distances between pairs. On the other hand, feature-based methods typically use a sum of distances between paired features to evaluate an alignment. When MS/MS identifications are available, they can be incorporated into the score function to penalize alignment paths that deviate too far from corresponding peptide features.

To increase the robustness of an alignment method, it is often necessary to allow and deal with mismatches [WW05, KSS+07, FGR+06]. This is particularly important in biomarker discovery, where one explicitly looks for regions of non-similar signals that might distinguish between different samples. For general profile alignment, Prince and Marcotte [PM06] note that gap penalties allow ignoring noise through big gaps given enough evidence, and even linear gap penalties yield better alignment performance; non-linear penalties did not appear to improve the alignment. Less robust approaches that do not take care of mismatches may still be able to yield good results, as, in most samples, persistent signals from expressed housekeeping genes provide sufficient information for computing the alignment [PMW+06, LNRE05, PM06].

6.4.2.4 Optimization Algorithms

Most alignment algorithms optimize the alignment with regard to their score function. If the elution time axis is discrete, most methods introduce the assumption that elution order of peptides is conserved throughout experiments. This allows the use of dynamic programming approaches, which are guaranteed to efficiently compute the global optimum. If, on the other hand, the time axis is continuous, the search space becomes infinite and one may only find a reasonable approximation to the optimum. Gradient descent methods can then be used as a general optimization strategy. However, they are sensitive to local minima and may thus become inefficient.

More specific optimization strategies are based on additional model assumptions. For example, the generative model of Listgarten et al. [LNRE05] is optimized using the Baum-Welch algorithm, which is a specific version of the EM algorithm [BPSW70] for HMM models.

Petritis et al. [PKF+03] use a genetic algorithm to explore the search space for aligning multiple images. Such a procedure does not guarantee to find the optimal solution, and only a small part of the search space is actually explored. However, algorithms that guarantee to find a global optimum are often infeasible for complex search spaces.

6.4.2.5 Choice of a Reference

In contrast to pairwise LC-MS image alignment, aligning multiple images becomes infeasible even for small numbers of images, since the computational demand grows exponentially. Similar to multiple sequence alignment, most multiple alignment approaches work by establishing a reference image. Here, a reference means either a selected experiment, or a model against which remaining experiments are progressively aligned. To our knowledge, only a few methods propose multiple alignments without using a reference image [PKF+03].

Some approaches establish a static reference model image to which all single images are aligned [JMP+06, NMA+05, LNRE05, WTF+07, LGST+07, ZAA+05]. Such a model image can be built from the current set of LC-MS images [LNRE05, WTF+07] by using robust methods, such as the median [ZAA+05]. It can also be built from previous experimental data such as accurate mass time tag (AMT) databases [JMP+06, NMA+05]. If the model is kept static throughout the alignment process, all images are aligned relative to this model.

Other approaches do not explicitly build a model, but use an image from the used dataset instead as a reference image [RJR+04, Eil04, PMW+06, KTZ+06, KSS+07, KO05, PM06, FGR+06]. Possible choices for a reference experiment include the first [Eil04, ACvG+06], the cleanest [ZAA+05], the most complete [TvdBA04, LGST+07], or the most representative [KTZ+06, BDMM02] experiment in the dataset. Obviously, any of these choices can lead to potential biases.

Yet other approaches do not keep a static reference image, but rather recompute the reference with each pairwise alignment. Listgarten et al. [LNRE05] describe a multiple alignment that is obtained by pairwise alignments using a probabilistic hidden Markov model. Likewise, Mueller et al. [MRS+07] create the reference image by hierarchical clustering of the data.

Even in a few pairwise alignment methods, a reference needs to be chosen, and the result of the alignment depends on that choice. Pairwise alignment methods typically treat both datasets equally, but some papers also propose asymmetric alignments. Fischer et al. [FGR+06] perform an evaluation of the consistency of such alignment methods.

6.4.3 Inspection and Validation of Elution Time Alignments

In large-scale experiments, direct validation of elution time alignment approaches is difficult, as independent "ground truth" information is frequently not available. Nevertheless, an indirect validation is often possible by either evaluating the alignment efficiency in a larger application context using visual inspection, by comparing an approach to other methods, or by using available small-scale ground truth data.

6.4.3.1 Visual Evaluation

The quality of an alignment can often be estimated using visual evaluation, where one usually compares the two LCMS images by plotting them side by side (Fig. 6.8) or by overlaying the TIC profiles or LC-MS images [RJR+04, PMW+06]. Displacement graphs [ACvG+06, KTZ+06, BDMM02] plot elution time shifts against the elution time (see Fig. 6.10 for a variant). Silva et al. [SDD+05] use this graph to actually compute the alignment itself. In dynamic programming approaches, heat maps of the alignment matrix are used to assess the quality of the optimal alignment compared to other possible paths (cf. Fig. 6.11).

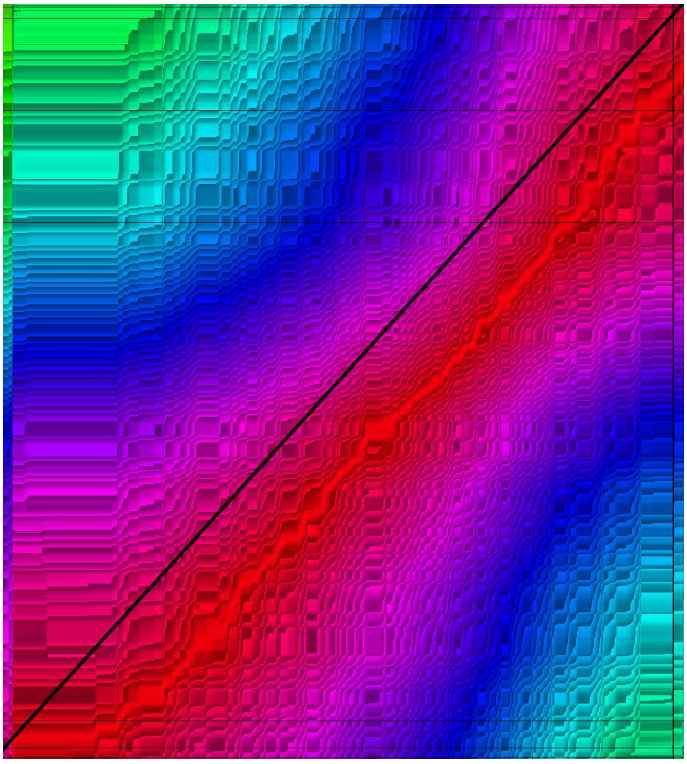

Figure 6.11: Heat map of an alignment matrix as computed by CHAMS [PMW+06]. Red colors represent regions of high alignment score, whereas regions of low score are given in blue and green. The optimal alignment path falls into the bright red regions, of maximal score. Larger regions of similar color indicate image parts of low signal; an alignment path has less confidence in those regions. As noted in Fig. 6.10, the bright red regions suggest a non-linear elution time distortion. Heat maps are a tool to visually assess alignment quality (cf. Section 6.4.3).

6.4.3.2 Validation using Ground Truth

In GC-MS, TIC landmarks tend to be easier to identify than in LC-MS, since elution ranges are typically smaller. It is thus possible to generate a benchmark dataset. For example, van Nederkassel et al. [vNXL+06] identify ten peaks manually and monitor their chromatographic coordinates in different temperature conditions, before and after alignment.

In many cases, pairs of features that should be aligned with each other are known beforehand [KSS+07]. Assessing the quality of an alignment with regard to these features is also easy, for example by using a distance between corresponding features [PM06]. Wang et al. [WTF+07] count the number of expected and identified peptides in an experiment that includes a spiked peptide.

Kirchner et al. [KSS+07] and Bylund et al. [BDMM02] validate their approaches establishing a ground truth by applying a known LC elution time distortion to the data. Although this allows evaluating the accuracy of alignments, it does not test their robustness. Accuracy of an alignment can also be measured using the percentage of matched peaks and the standard deviation of the elution times across the aligned LC-MS images before and after warping, given a set of known reference peaks. This method is also well suited for parameter optimization of some algorithms.

6.4.3.3 Validation in the Absence of Ground Truth

In the absence of available ground truth, most approaches have to rely on first establishing a reasonable substitute. Several papers use correlation to assess the quality of the alignment, for example TIC correlation [vNXL+06, WW05, KTZ+06], or the correlation of matched peptide profiles [WTF+07]. One substitute uses pairs of previously acquired MS/MS identifications.

Prakash et al. [PMW+06] compare alignment-based protein identification to MS/MS identifications of the same features and quantify the overlap. Prince and Marcotte [PM06] also use MS/ MS identifications as assumed ground truth. They discard peptides that are sequenced and identified more than twice as these tend to correspond to broad, imprecise LC peaks. Alignment accuracy is then defined as the distance between aligned MS/MS. It should be noted, however, that MS/MS identifications could be problematic as the only basis for alignment. Peptides are not consistently selected for fragmentation, due to the low sampling rates of the LC-MS process, and they may be selected at different times during the elution of the peptide (see Fig. 6.8).

LC-MS image features, as opposed to MS/MS spectra, are typically not characteristic enough to reliably identify a peptide. Nevertheless, using an AMT-like identification approach [JMP+06, PMW+06], such features can substitute ground truth.

Finally, one may use downstream analyses to afterwards estimate the accuracy of an alignment. As an example, changes in predictive power for biomarker discovery [NMA+05, ACvG+06, LNRE05] before and after alignment can be measured using principal component analysis [BDMM02]. Radulovic et al. [RJR+04] evaluate the performance of the whole platform in different proteomics application scenarios.

6.4.4 How to choose an alignment method

As LC-MS images can represent significant amounts of data, their handling and analysis is a nontrivial, but essential step with an important impact on the quality of the downstream analysis results. The status of the field – in particular, regarding validation – is not advanced enough to allow the compilation of general guidelines for the choice of an alignment algorithm. However, two general observations may narrow the choice in a given application scenario.

First, finding linear transformations is much faster and more robust than finding a nonlinear transformation. Therefore, we recommend performing an initial assessment of non-linear behavior in the data, using fully general software that allows nonlinear transformations. In a second step, a method that matches the observed complexity ("no transformation necessary", "linear transform" or "non-linear transform") can be chosen accordingly.

The second consideration is that the reviewed methods are not always applicable, depending on whether the data have been acquired or not.

1. For instance, some methods require prior MS/MS identifications, and are thus not applicable to LC-MS data without MS/MS identifications.
2. Methods that are designed for use with profile data should be applied with caution on data recorded in data reduction mode with the manufacturer's software. Centroiding and noise reduction are likely to affect the alignment results. At a minimum, it should be ensured that the same parameters for data acquisition are used for all experiments.
3. Certain methods – in particular, profile-based alignment – may require significant computational resources in terms of compute cycles and/or memory, especially for multiple alignments.

4. On the other hand, feature-based methods require a data extraction step that may be error-prone, particularly for low-resolution data.
5. As of the writing of this review, software availability is also an issue. Some programs are not available for all platforms (Windows, Linux, etc.) or require a specific programming environment (e.g. Python, R, Matlab). Only few methods are publicly accessible on web servers, and many provide only sparse documentation.

In addition to the collections of software given in Katajamaa et al. [KO07], Palagi et al. [PHWA06] and Codrea et al. [CJHM07], Section 6.1 gives an overview of available software that are specifically targeted to align multiple LC-MS images.

Software	Description	Type	URL	Language
OpenMS	Library for the analysis, reduction and visualization of LC-MS (/MS) data	Platform and pipeline	http://open-ms.sourceforge.net/	C++
TOPP	OpenMS protein identification / quantitation pipeline	Platform and pipeline	http://open-ms.sourceforge.net/TOPP	C++
TPP	Institute for Systems Biology "Trans-Proteomic Pipeline"	Platform and pipeline	http://tools.proteomecenter.org/software.php	C++, perl
XCMS	Software package for metabolite profiling from LC-MS data	Platform and pipeline	http://metlin.scripps.edu/download/	R
CPM	Alignment of time-series data (as in LC-MS[/MS]) using Continuous Profile Models	Calibration / alignment	http://www.cs.toronto.edu/jenn/CPM/	MATLAB
OBI-Warp	Aligns multiple LC-MS (/MS) datasets in elution time by Dynamic time warping	Calibration / alignment	http://obi-warp.sourceforge.net/	C++
Recalibrate_ using_MSMS	Recalibrate LC-FTICR MS (/MS) datasets using peptides identified by MS/MS	Calibration / alignment	http: //ms-utils.org/recalibrate_using_MSMS.html	C
SpecArray	Aligns multiple LC-MS datasets and much more	Calibration / alignment	http: //tools.proteomecenter.org/SpecArray.php	C++
ChAMS	Match all signals from two LC-MS experiments by alignment. Alignment is done as to maximize the overall pair wise similarity between matched MS spectra.	Alignment	http://www.pasteur.fr/recherche/unites/ Biolsys/chams/index.htm	C
LCMSWARP	Elution time alignment and feature clustering	Alignment	http://ncrr.pnl.gov/software	C++
PETAL	Peptide Element Alignment for LC-MS data	Alignment	http://peiwang.fhcrc.org/researchproject.html	R
SuperHirn	Peak detection, alignment and normalization.	Alignment	http://tools.proteomecenter.org/wiki/index. php?title=Software:SuperHirn	C++

Table 6.1: Software packages for aligning LC-MS images.

6.5 Retention Time Alignment (Conclusion)

Global LC-MS-based profiling technology is increasingly employed to elucidate complex cellular processes and their correlated phenotypes. Because of unavoidable experimental variation, the high dimensionality of experimental data, and finite resolution and accuracy of the instruments, computational tools are necessary to transform these datasets into reliable and interpretable information on proteins. Scientifically validated data preprocessing tools, among them, the LC-MS image alignment algorithms surveyed here, are critical for any application, such as biomarker discovery and protein identification from multiple LC-MS experiments.

Entirely satisfactory comparison and validation of the currently available alignment methods remain an important but elusive goal, mainly because of the absence of a ground truth in large-scale LC-MS analyses. The assessment of alignment quality is challenging too, because the information contained in MS/MS spectra in many cases still appears to be significantly richer, but less dense, than even highly accurate information on mass and elution time.

Image alignment is a major new approach to circumvent the limitations of MS/MS-based analysis, and to exploit the full potential of MS data, moving towards higher proteomic coverage. Even though the propagation of peptide identity remains difficult, even with highly reproducible chromatography, an appropriate alignment and identification strategy tuned to the available experimental data can confidently identify many peptides and proteins, and enable new experimental strategies. The avoidance of feature detection before LC-MS image alignment significantly improves the sensitivity of the approach.

The existing and emerging computational algorithms, along with their systematic evaluation, will eventually help to underpin the full biological potential of the technology. Its success is likely to be driven by parallel and complementary improvements of the experimental LC-MS platform itself; in particular, concerning improvements of its effective dynamic range. Future developments in computing and statistics in LC-MS data preprocessing can be expected to take full advantage of the lessons learned from the analysis of large-scale gene expression microarray data. In particular, small sample sizes are likely to result in spurious biomarker discoveries.

With a growing number of publicly accessible LC-MS data repositories, alignment remains a key tool to increase the statistical power of experimental observations, and to leverage data and information already available from different experiments and laboratories. Such integrative analysis procedures will critically rely upon effective data preprocessing that can deal with numerous sources of both experimental and biological variability, across a variety of biological samples, and across different experimental conditions, and if possible, laboratories. Although careful control of the experimental protocol, including flow rate control, nano-LC, air conditioning, better columns and automated robots, can effectively remove some of the unwanted variation, there is still a great need for complementary future improvements in the computational pre- and post-processing algorithms for LC-MS images.

As alignment methodology matures, new applications of alignment are being explored, such as the systematic quality assessment of large sets of experiments [PPW+07]. The strong increase in proteomic coverage possible in the case of many LC-MS experiments is starting to enable interesting new applications in systems biology [FRS+07]. We expect that, as modern approaches to biology can start to rely more and more on large sets of experiments, alignment will take an increasingly important role in the data analysis machinery of systems biology.

We gratefully acknowledge Tina Guina and David R. Goodlett for making available to us the data underlying Figs. 6.8, 6.9, 6.10 and 6.11 [GRB+07], Amol Prakash for providing us with the ChAMS [PMW+06] source code, and the anonymous referees for their careful reading and constructive feedback on this manuscript.

6.6 M/z calibration

A mass spectrometer is a very accurate instrument, and like all accurate instruments, it needs careful calibration procedures to unleash its potential. When coupling a mass spectrometer to

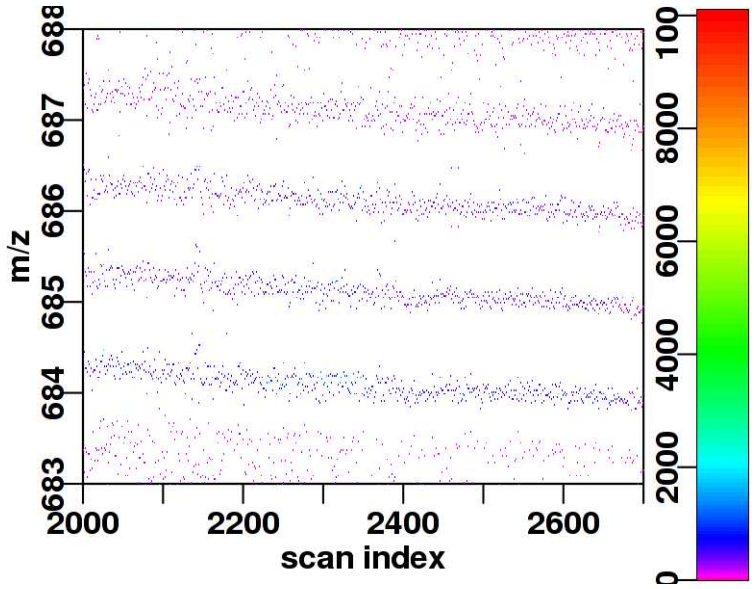

Figure 6.12: A mass spectrometer can lose its m/z calibration over time during the course of LC/MS analysis. The figure was generated from data acquired on a LTQ instrument (linear ion trap).

liquid chromatography, this is a stronger issue because the mass spectrometer may lose its calibration over time. In this section, we only briefly present m/z calibration. For more details we refer the reader to [ZM07, Mat07].

6.6.1 Experimental reasons for calibration

Over the course of the experiment, subtle changes in the setup can alter the accuracy of the mass spectrometer. Figure 6.12 shows m/z drift in the later stages of the LC/MS experiment. Here are several examples of experimental reasons for calibration problems.

1. TOF analyzers are sensitive to temperature changes because the power supply output and the length of the flight tube are functions of the temperature[CLT01]. Although calibrated analyzers may have an accuracy up to 10 ppm, the power supply can induce variations of 25 ppm per degree Celsius, and stainless steel has a coefficient of expansion of about 18 ppm per degree Celsius.
2. Maldi-TOF instruments have additional calibration issues because calibration is dependent on the sample position on the MALDI plate [MCP$^+$03].
3. In FT-ICR instruments, ions do not have exactly circular trajectories. In that case, the Fourier transform does not yield accurate measure of the rotational frequency and of the m/z ratio [SRS03].

6.6.2 Technical solutions for calibration

External calibration To calibrate a mass spectrometer, the standard procedure is to introduce a calibration compound in the mass analyzer and compare the recorded m/z values with the theoretical values. This calibration reference can be introduced at regular intervals during the LC/MS analysis for measuring calibration scans. [Cha03] gives a technical solution to inject the calibration compound between the LC column and electrospray ionization, without modification of the mass spectrometer.

Internal calibration When using internal calibration, the reference compound is usually mixed with protein sample. This ensures that the lock mass for calibration is available in all mass spectra and that it is not separated by liquid chromatography. However, the lock mass signal may hide other low-intensity signals in the biological sample because of competitive ionization (see Section 2.3. Known contaminants can be used as internal calibration standards.

MS/MS calibration After identification of MS/MS spectra, we can compute the theoretical m/z value of the parent ion, and compare it with the measured m/z value.

6.6.3 Computational approaches for calibration

M/z calibration is a similar problem to retention time alignment, but easier. It usually suffices to find an affine transformation of the m/z axis, and only two mass measures are required for this. For example, in TOF instruments, calibration consists in finding the coefficients A and B such that

$$t = A\sqrt{m/z} + B$$

Some computational approaches have been proposed for mass calibration without using a reference compound. In [WLJR05] more non-linear calibration functions are used to improve m/z ratios throughout the m/z range. [BM08] addresses the problem of peptides in the sample having nearly the same m/z ratio as the reference compound, and provides a robust method.

Conclusion

In general, current calibration procedures are based on a reference standard and yield adequate results. Global calibration with affine transforms are sufficient in most cases.

6.7 Intensity baseline

Preprocessing methods that correct the intensity values in LC/MS images address the following problems:

- baseline correction removes the contribution of chemical noise to area under the curve quantification,
- denoising smooths the intensity function,
- normalization corrects global and artificial differences in protein signal intensity.

Baseline correction is presented in this section, and normalization in Section 6.8. The noise model presented in Chapter 7 can also be used for baseline removal and normalization.
Denoising of mass spectra is not presented in details in this manuscript, because the methods employed are not specific to LC/MS images. Standard signal processing tools are applied to the data such as smoothing, wavelet filters, or morphological filters. For more information on denoising for LC/MS images, we refer the reader to [LNRE05, Mat07, SDM$^+$04, YHY09]

6.7.1 Decomposition of noise into baseline and residuals

The intensity function $\mathcal{I}(t, m)$ measured on the LC/MS platform is decomposed into two independent components:

- peptide signal $\mathcal{S}(t, m)$ is a deterministic function of the retention time and m/z ratio,
- background noise $\mathcal{N}(t, m)$ is the realization of a non-stationary random process.

Peptide signal is modeled in more detail in Chapter 8. In this section and the next, we study the background noise component.

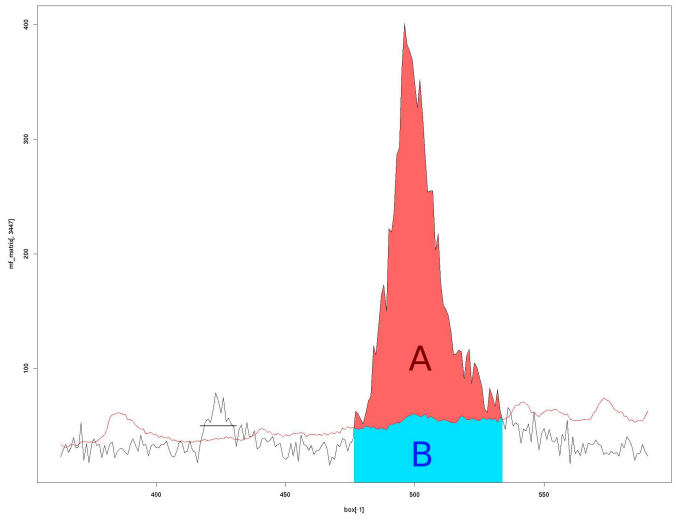

Figure 6.13: Area under the curve quantification. The baseline (red curve) is estimated as indicated in Section 6.8. **A** corresponds to the area under the curve attributed to the peptide signal. **B** is the contribution of the baseline.

The background noise process is itself decomposed into a deterministic component $B(t, m)$ called *baseline* and residuals $\varepsilon(t, m)$ such that:

$$B(t, m) = \mathbb{E}[\mathcal{N}(t, m)]$$
$$\varepsilon(t, m) = \mathcal{N}(t, m) - B(t, m)$$

From the point of view of time series, $B(t, m)$ contains both the trend and the seasonal variations. In some papers, only the residuals $\varepsilon(t, m)$ are designated as "noise".

The decomposition of intensity into baseline and residuals was first presented for the analysis of isolated mass spectra. The nomenclature has been extended to LC/MS images, even though it would be more accurate to say "base surface" in two dimensional separations.

Baseline is especially important in MALDI ionization because of the formation of matrix clusters [SDM+04, MPP+07]. These create patterns similar to contaminants, see Figure 6.5 on 86.

6.7.2 Consequences for quantification

As explained in Section 4.5, quantitative information in LC/MS images is measured by the area under the curve. As indicated on Figure 6.13, the area under curve of a peptide signal can be decomposed into:

$$\int \mathcal{I}(t, m) = \underbrace{\int \mathcal{S}(t, m)}_{A} + \underbrace{\int B(t, m)}_{B} + \underbrace{\int \varepsilon(t, m)}_{\simeq 0}$$

The contribution of the baseline for quantification of low intensity peptides is significant. In Figure 6.13 it represents more than 20% of the area under the curve. The contribution of the residuals to the area under the curve is a random variable with null expectation.

For relative quantification and biomarker discovery, the baseline contribution affects the computation of intensity ratios $\mathcal{I}_1/\mathcal{I}_2$ for peptides in different LC/MS images because the baseline is not the same (see Figure 6.14 on page 113 for an example) :

$$\frac{\mathcal{I}_1}{\mathcal{I}_2} = \frac{\int \mathcal{S}_1(t,m) + \int B_1(t,m)}{\int \mathcal{S}_2(t,m) + \int B_2(t,m)} \neq \frac{\int \mathcal{S}_1(t,m)}{\int \mathcal{S}_2(t,m)}$$

Remark Baseline has a marginal effect on peptide identification. In MS/MS identification, parent ions are selected for fragmentation based on their intensity in the parent MS scan. However, the selected parent ions are very intense in general, and the baseline contribution is low. For MS identification, the main difficulty is in feature detection as presented in Chapter 8.

6.7.3 Baseline removal

Preprocessing methods deal with baseline by first estimating the baseline function $B(t,m)$ then subtracting it from the intensity function $\mathcal{I}(t,m)$. Further analyses are carried out on $\mathcal{I} - B$ with the underlying hypothesis that baseline removal is perfect and that noise is centered (null expectation).

Baseline removal is usually performed in each mass spectrum separately, from a single observation of the noise process $\mathcal{N}(t,m)$. To estimate the mean intensity of the noise $B(t,m)$, baseline estimation relies on the hypothesis that the function $m \mapsto B(t,m)$ is sufficiently smooth. Conversely, the peptide signal is considered as a series of pulses, in contrast to slow variations of the baseline level.

Most baseline estimation approaches fall into two categories:

1. Start from local minima of the intensity, then build a smooth baseline by linear interpolation [CTM+05, MPP+07, YPT+03, BCF+06, LGL+05, KHS+07], moving average [DSPA07], finding the convex hull [LKP+03], etc.
2. Use robust estimates of the noise location like local linear regression [WNP03], polynomial regression [WCD+05], LOESS regression [LGL+05, WNP03], wavelet filters [DKL06, LGR+06, SS04], ...

In MALDI spectra, the baseline function has been modeled as a decreasing exponential [WNP03].

We are not aware of baseline estimation procedures that deal with baseline estimation in two dimensions. One such approach is presented in Section 6.8 as a side effect of normalization and another in Chapter 7 as a consequence of the noise model. They have not been sufficiently validated for practical application to LC/MS data. Neighboring mass spectra provide repeated observations of the noise process, which should in principle facilitate the estimation of the mean noise intensity. However, peptide signals are not impulsional in the chromatography dimension, so much so that the smoothness hypothesis in the m/z axis is still required.

The smoothness hypothesis used for baseline estimation ignores an important aspect of the background noise: its pseudo-period. In all the LC/MS images generated in the course of this PhD thesis (see Figure 6.14 on page 113 for example), we have observed that chemical noise is variable in the m/z direction, and that is has a 1 Da period. The baseline should have a corresponding pseudo-period of 1 Da. More generally, we expect the baseline function to contain features of low frequency in the sense of Fourier analysis or large scale in the sense of scale space theory, but up to 2 Hz, i.e. two oscillations per Da. At this scale, peptide signals are not impulsional because the mass analyzer does not have enough resolution. At 10,000 resolution, the peak width of peptide signals is roughly 0.1 Da for ions of 1,000 Da.

Remark In the feature detection method proposed in Chapter 8, baseline estimation and removal are not necessary. Consequently, we suggest not to use it to preserve the statistics of the background noise intensity.

6.8 Intensity normalization

Intensity normalization or standardization corrects systematically high or systematically low intensity values related to variations in the sample preparation. Most of the observations and statements presented in this section is original work, and has not been previously reported in the literature to our knowledge.

As normalization changes the intensity values in the LC/MS image, it has a corrective effect on quantification, and in turn, biomarker discovery. In absolute quantification (described in Section 4.5.2), unnormalized intensity values indicate the real protein concentration in the analyzed sample, but not the intended concentration, which is obtained after division by the normalization factor. In relative quantification (described in Section 4.5.3), isotope labeling methods are not affected by normalization problems, but label-free methods are.

6.8.1 Experimental reasons for normalization

Normalization corrects for variations in the sample concentration:

- dilution steps divide the concentrations of all proteins,
- if a biological is kept as dessicated powder, it needs to be dissolved and diluted prior to analysis on a LC/MS platform,
- when analyzing cell lysates, the number of selected cells may be variable.

These experimental variations account for a global multiplicative factor, called the *normalization factor*, that is applied to the concentration of all proteins in the sample, and hence to all the intensity values of signals. When comparing the LC/MS images of repeated experiments, the global normalization factor can be observed as systematic intensity differences between the two images as in Figure 6.14.

Other sources of experimental variation account for normalization factors for individual mass spectra in the LC/MS image:

- saturation in the LC column may reduce the effective flow rate. Consequently, a given mass spectrum may correspond to decreasing amounts of material.
- if there is saturation in the ion source, the peptides in the sample compete for the available charges. We may obtain the same distribution of intensity as a function of m/z ratio, but the intensity values will be downscaled.
- even without saturation, competitive ionization can lead to the reduction of all signals other than the most intense ion.
- "automatic gain control" is used to optimally fill ion traps and FT-ICR traps. It is unclear whether other types of instruments implement some level of "automatic gain control".
- in electrospray ionization, the Taylor cone may be unstable. As the peptide ions enter the mass analyzer through a tiny hole, fluctuations of the Taylor cone may lead to sample loss.

These reasons call for spectrum-specific normalization factors. The gain factor can change abruptly from one scan to the next, thus producing visible changes in the overall intensity of a mass spectrum. These problems can be observed in a LC/MS image as vertical stripes, as shown on Figure 6.15. Most of the datasets that we have encountered suffer from these vertical stripes, regardless of instrument type.

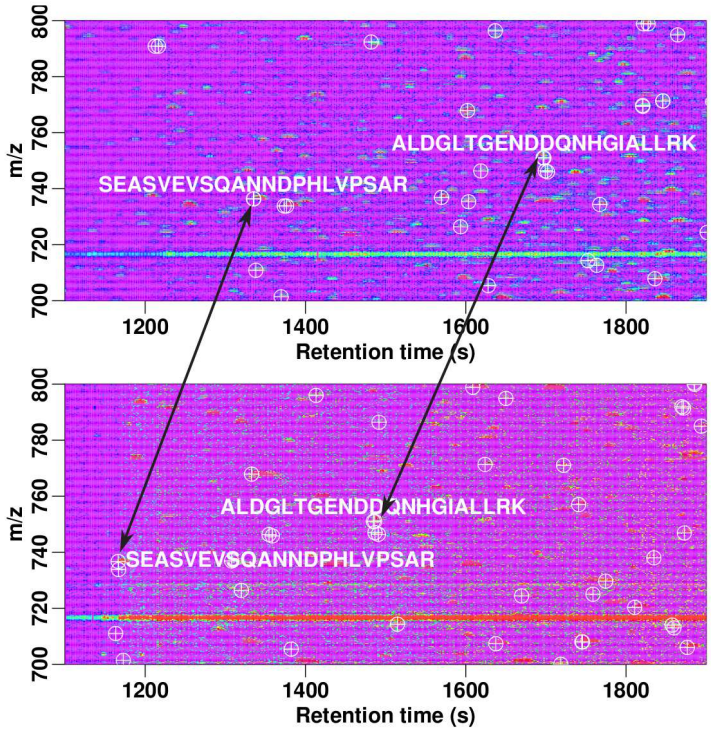

Figure 6.14: LC/MS images of repeated experiments show systematic differences in the signal intensity related to global normalization problems.

Spectrum normalization affects quantification in the measured intensities of each peptide signal. It also has an indirect effect on the correction of retention times. Intensity-based methods such as ChAMS (Section 6.2) and others reviewed in Section 6.4 match mass spectra of similar intensity and are likely to align the vertical stripes. This is a feature if the normalization factors are reproducible and can be compared between experiments, it is a source of error otherwise.

The need for a global normalization factor related to dilution is well established, but not the need for spectrum specific normalization factors. To our knowledge, this has been reported only in [LNRE05] with a Hidden Markov Model containing spectrum-specific gain parameters. That algorithm was originally designed for processing of speech signals where the volume may be time dependent, and the gain parameters have not been motivated in the LC/MS context.

Normalization problems affect both peptide signal intensities and (chemical) background noise as shown in the examples in Figure 6.14 and 6.15. The approaches described in the rest of the chapter try to match the peptide signal levels, the background noise levels, or both.

6.8.2 A classification of normalization problems

Let $\mathcal{I}(t, m)$ denote the intensity values in the LC/MS image as a function of retention time t and m/z ratio m. In spectrum normalization, t and m have different roles. We write $\mathcal{I}_t(m)$ to emphasize the fact that the intensity values are organized in MS spectra and that normalization is spectrum specific. We denote NF_t the normalization factor, as a function of the retention time or scan index t.

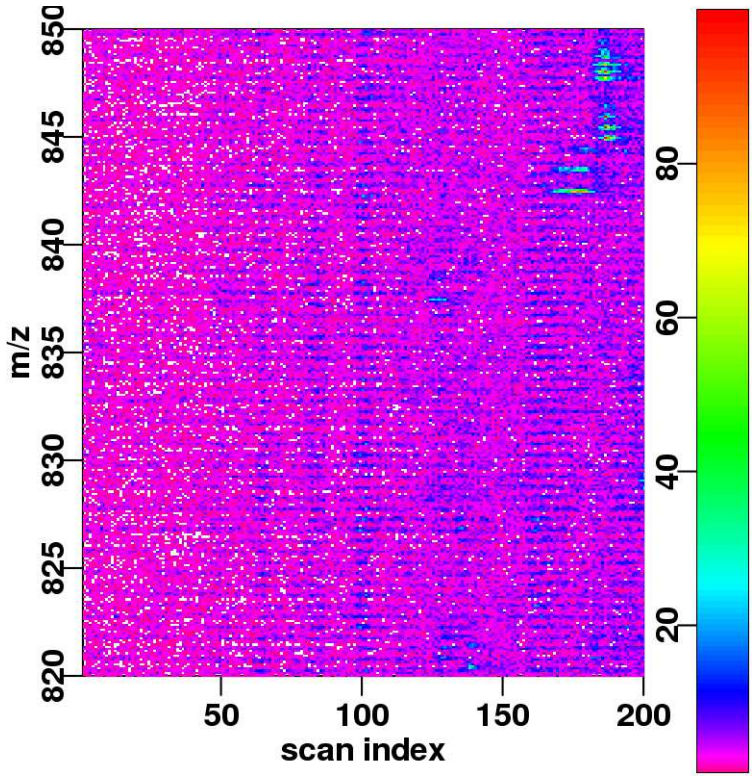

Figure 6.15: LC/MS images obtained on a Q-TOF instrument exhibit spectrum normalization problems. These are recognized as vertical stripes in the image.

Depending on the hypotheses on $\mathcal{I}$ and NF, normalization problems fall into the following categories.

Problem 1 (normalization with a reference): Intensity is modeled as:

$$\mathcal{I} = \mathrm{NF}\,(\mathcal{N} + \mathcal{S})$$

where $\mathcal{N}$ and $\mathcal{S}$ represent background noise and peptide signal intensity and are independent identically distributed random variables. The distribution of $\mathcal{N}$ or $\mathcal{S}$ are known from the reference image. $\mathcal{I}$ is not dependent on retention time or m/z ratio; this model can be applied to global normalization or spectrum normalization.

Problem 2 (multiple normalization): Intensity is modeled as:

$$\mathcal{I}_t = \mathrm{NF}_t\,(\mathcal{N} + \mathcal{S})$$

In this model t can denote retention time for spectrum normalization, or LC/MS image index for global normalization. Without a reference image, the absolute intensity level cannot be determined for absolute quantification.

Problem 3 (normalization with a spatial dimension): In mass spectra, the noise level is dependent on the m/z ratio, and cannot be considered as identically distributed:

$$\mathcal{I}_t(m) = \mathrm{NF}_t\,(\mathcal{N}(m) + \mathcal{S}(m))$$

where m is the m/z ratio. This problem takes heteroskedasticity[4] into account. When performing normalization based on lists of peptides, m can be interpreted as the peptide index.

Problem 4 (normalization of spatial processes): After calibration, the sampling grid may be different in mass spectra and may require interpolation:

$$\mathcal{I}_t(m) = \mathrm{NF}_t\left(\mathcal{N}(m) + \mathcal{S}(m)\right)$$

When using interpolation or other resampling methods, this problem reverts to Problem 3.

Problem 5 (real data): Spectrum specific normalization in real data combines the previous problems, with the additional difficulty that the noise distribution is spectrum specific (mostly due to the baseline component):

$$\mathcal{I}_t(m) = \mathrm{NF}_t\left(\mathcal{N}(t,m) + \mathcal{S}(t,m)\right)$$

This is an ill-posed problem: when intensity is higher, is it because of normalization, baseline, or cancer[5] ?

Problem 6 (multimodal normalization): Normalization is used in label-free quantification methods to compare the intensities of several *different* experiments. This corresponds to:

$$\mathcal{I}_t^k(m) = \mathrm{NF}_t^k\left(\mathcal{N}^k(t,m) + \mathcal{S}^k(t,m)\right)$$

where k represents different image modalities (disease and control experiments, time points, species, etc.)

Related problems Normalization is an instance of image registration where no spatial transformation is necessary but only standardization of the image intensity. In image registration, most approaches focus on finding the adequate spatial transformation. In contrast, normalization focuses on the intensity (gray level) and tries to preserve quantitative information. Spatial transformations are performed by retention time alignment and m/z calibration.

In the analysis of gene expression profiles with microarrays, normalization corrects for systematic differences of detector gain (variations of hybridization efficiency). Some methods deal with spatial variations of the normalization factor across the array. Other methods deal with overlapping contributions from the different dyes. Additional details can be found in [YDL+02]

6.8.3 State of the art in intensity normalization

Applying a global normalization factor consists in multiplying all the intensity values in one experiment to match to another experiment. In most approaches, the total protein content of the LC/MS image, as represented by the total intensity (total ion count), is made to agree between LC/MS images [BMW03, WNP03]. Variations on the total ion count include the mean peak intensity [SS04].

[THN+04] use the quantiles of the intensity distribution to standardize the intensity values in a mass spectrum. They use an affine transformation to map the 10th and 90th percentiles to 0 and 1 respectively.

[4] Non uniform variance.

[5] Some tumors lead to elevated levels of most proteins, as discussed in Chapter 5

As normalization builds on the hypothesis that the concentration of "housekeeping" proteins is the same, some methods try to ensure that the intensities of a list of peptides are comparable. These methods measure the center of mass of the intensity distribution like median peak intensity [WZL+03], or use the median ratio in pairs of peaks [ARL+04].

As only the most abundant proteins may have reproducible concentrations, [WTZ+06] consider only the L most intense signals and their median intensity, whereas [RJR+04] find the normalization factor such that only 100 signals have intensity over 10,000.

6.8.4 Normalization based on the background noise.

In Sections 6.8.5 and 6.8.6, we evaluate several indicators of the normalization factor based on analyzing the background noise and considering peptide signals as outliers. This is based on the hypothesis that normalization affects chemical noise in LC/MS images.

Using the background noise for normalization is against the current trends in the literature. Current methods rely on the hypothesis that the peptide content is the same across samples. Normalization is improved by removing the noise contribution to intensity values (baseline removal and denoising) and considering only the intensity from peptide signals.

Methods based on peptide signals fail to take advantage of the full information in LC/MS images. In particular, they use only the contributions of high-intensity signals. Performing normalization based on chemical noise has a few advantages. It uses the bulk of the dataset because although peptide signals are in large numbers, they are still isolated in LC/MS images. Moreover, we expect chemical noise to be useful in multimodal normalization (Problem 6), where the protein content varies in different LC/MS images. Finally, normalization of the noise intensity facilitates statistical analysis because the noise process may be considered as stationary. This assumption is used to estimate the noise level in Chapter 8 for the detection of peptide signals.

Validation Due to the lack of a suitable validation data set, we decided to evaluate the proposed spectrum normalization methods based on two criteria:

- The background noise distribution is the same in each mass spectrum. This is evaluated based on the stability of the quantiles of the intensity distribution in each mass spectrum, and the lower quantiles in particular.
- Some contaminants are not separated by liquid chromatography. They appear in each mass spectrum, and are expected to have a constant intensity throughout the LC separation.

The LC/MS images used in Section 6.8.5 and Section 6.8.6 were generated from the 4-5 protein mix described in [PMW+06]. The data file corresponds to a LC/MS analysis on a ESI-TOF mass spectrometer of a tryptic digest of 5 proteins (alodlase, catalase, a-lactalbumin, transferrin, b-lactoglobulin).

6.8.5 Spectrum normalization in isolated mass spectra

Problem statement For each mass spectrum with index t in the image (vertical column of pixels), we compute a normalization factor NF_t. After dividing the intensities of each mass spectrum with the normalization factor, the background noise distribution should be the same. This corresponds to Problem 1 in the proposed classification.

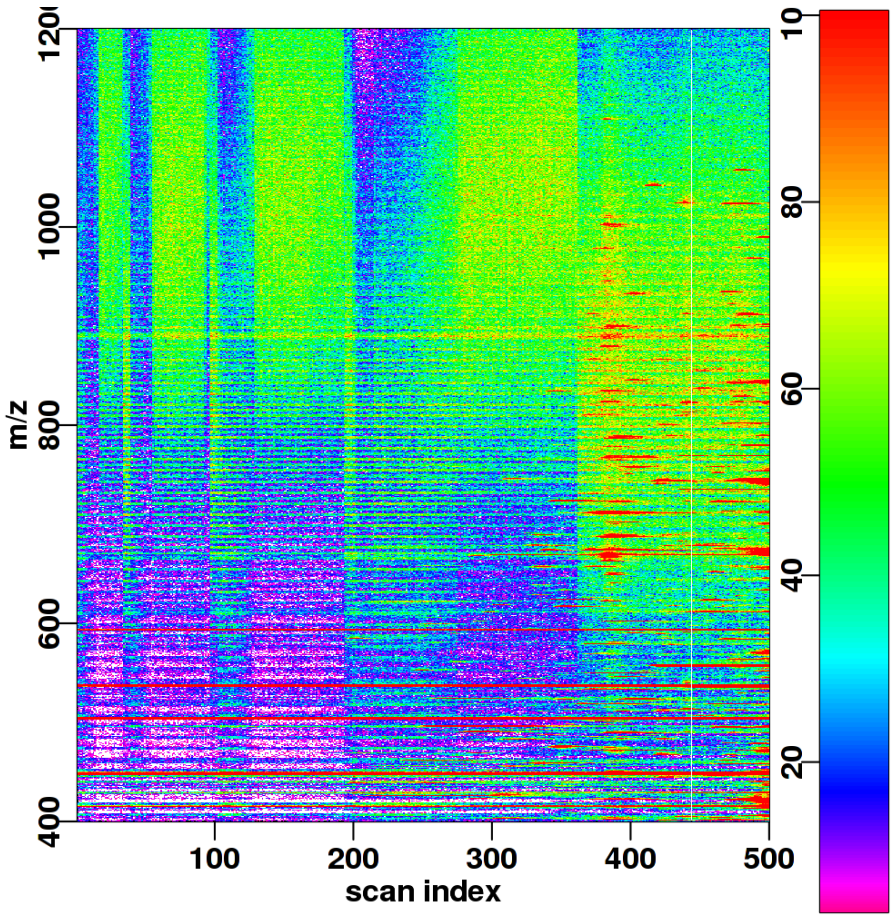

Figure 6.16: LC/MS image before normalization.

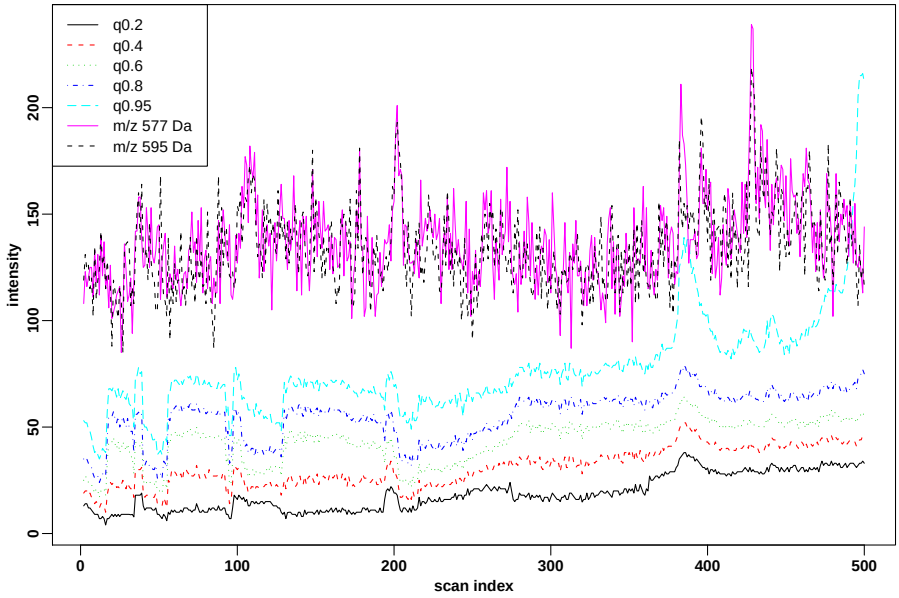

Figure 6.17: Quantiles of the intensity distribution and intensity of two contaminants before normalization.

In Figure 6.16, we show the LC/MS image before alignment, and the background noise distribution as a function of t, as indicated by a few quantiles in Figure 6.17. The curve corresponding to the quantile level 0.95 is expected to contain real peptide signals, and less emphasis is put on that level. We also check the elution profiles of two contaminants at m/z ratio 577 Da and 595 Da. After normalization, all the curves in Figure 6.17 are expected to be constants.

To compute the spectrum-specific normalization factors, we use the following measures of the location of the intensity distribution:

- total ion count in a mass spectrum (TIC, i.e. L^1 norm)
- mean intensity
- euclidean norm (i.e. L^2 norm)
- Huber m-estimator of location with the scale estimated with the median absolute deviation (MAD) [HWI81, VR02]
- median
- weighted mean of quantiles, we used $\mathrm{NF} = \frac{1}{3}\left(q_{0.3} + q_{0.5} + q_{0.7}\right)$
- contaminant intensity, we used $\mathrm{NF}_t = \frac{1}{4}\left(\mathcal{I}_t(577) + \mathcal{I}_t(578) + \mathcal{I}_t(595) + \mathcal{I}_t(596)\right)$

Huber m-estimator Let $(x_1, \ldots, x_n)$ denote a data set of real numbers. A m-estimator of location $\hat{\mu}$ is a solution to the following minimization problem:

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(x_i - \mu)$$

When $\rho(x) = x^2$, then $\hat{\mu}$ corresponds to the mean of $(x_1, \ldots, x_n)$. The Huber m-estimator corresponds to :

$$\rho(x) = \begin{cases} x^2 & \text{if } |x| \leq k \\ 2k|x| - k^2 & \text{if } |x| > k \end{cases}$$

The parameter k is adjusted by the user for a compromise between robustness (low values of k) and efficiency (large values of k). The standard choice is $k = 1.5s$ where $s = 1.4826\,\mathrm{MAD}$ is an estimate of the standard deviation based on the median absolute deviation.

Results Figures 6.18 to 6.24 show the LC/MS images obtained after normalization. Due to the non-uniform background noise in the chosen LC/MS image, it is not easy to evaluate the results, but the quantile plots are easier to interpret. These lead to the following remarks.

Before normalization, the intensity of the contaminants is more or less stable, as are the quantile functions. This supports the hypotheses for spectrum normalization.

Normalization with the total ion count is surprisingly good, as shown by the stability of the quantiles. However, this method fails in the presence of peptide signals around scan 500, and the contaminant intensities are disturbed.

Normalization with the mean is nearly identical to TIC normalization, although the number of sampling points may be different in successive mass spectra. Normalization problems seem less pronounced than in TIC normalization.

Normalization with the euclidean norm fails to provide stable noise background and contaminant intensity.

The Huber m-estimator is a robust version of the trimmed mean, and is less affected by the high intensity signals.

Normalization with the median is not convincing. In fact, in this data set, the median is useless because it is nearly always equal to 3, and normalization has no effect. Note that quantile regression [YLS03] cannot be applied in this context because we do not expect the normalization factor to be a smooth function.

Normalization with the weighted mean of three quantiles performs best, both in terms of the stability of the quantiles and of the contaminant intensity. It is not affected by the presence of high-intensity signals around scan index 500. In some places, normalization is off target because some of the considered quantiles are equal to 0.

For comparison, we normalize the LC/MS with the intensity of the contaminants, and ensure that the contaminant intensity is stable. It is not exactly constant because we used different contaminants in normalization and validation. Figure 6.24 shows the lack of agreement between contaminant intensity and background noise distribution. This raises doubts about the validation criteria.

In conclusion, normalization based on the quantiles is a promising method, but not completely satisfactory. Inconsistencies between the background noise quantiles and contaminant intensity prevent the validation of this approach (see the discussion in Section 6.8.7).

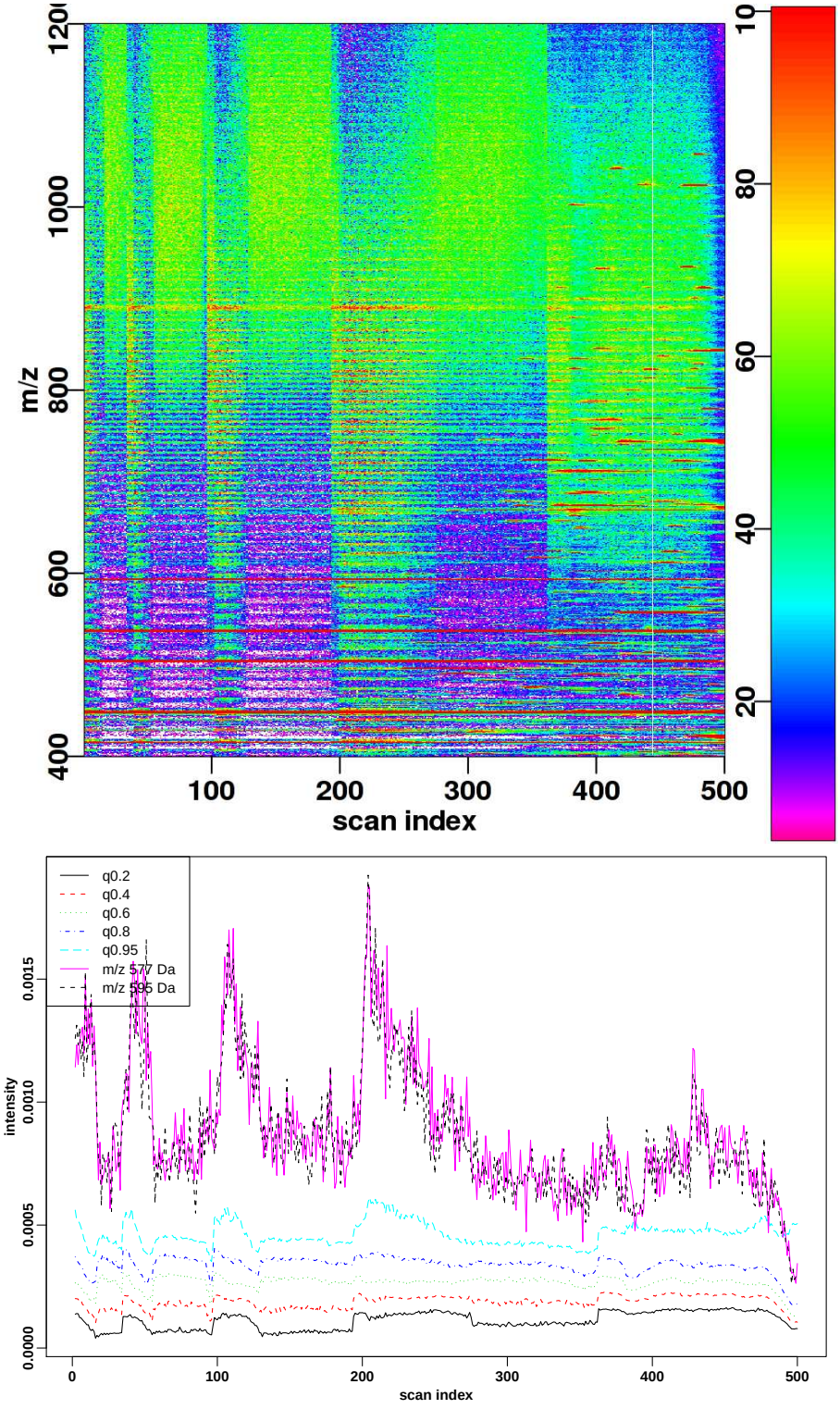

Figure 6.18: After TIC normalization.

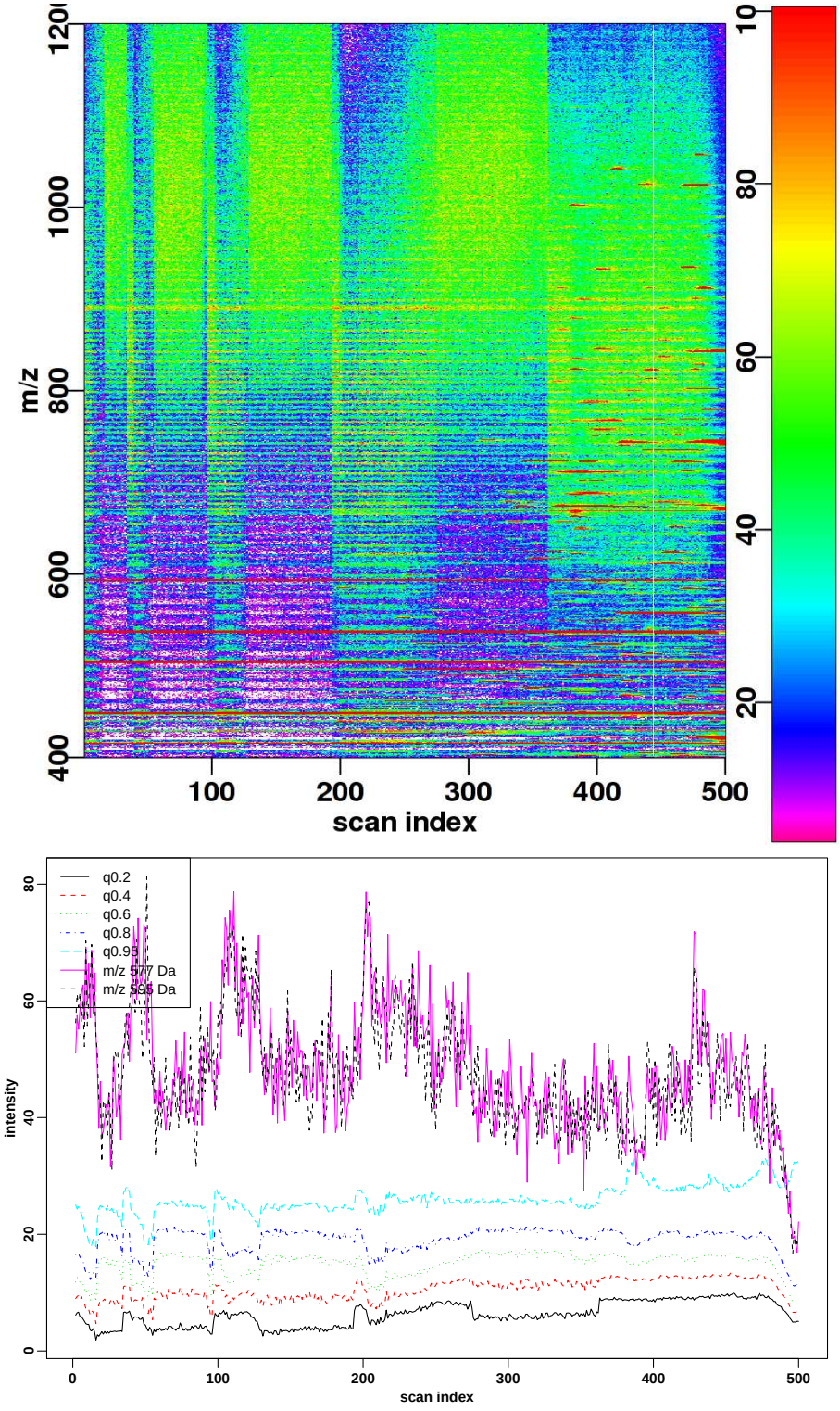

Figure 6.19: After normalization with the mean.

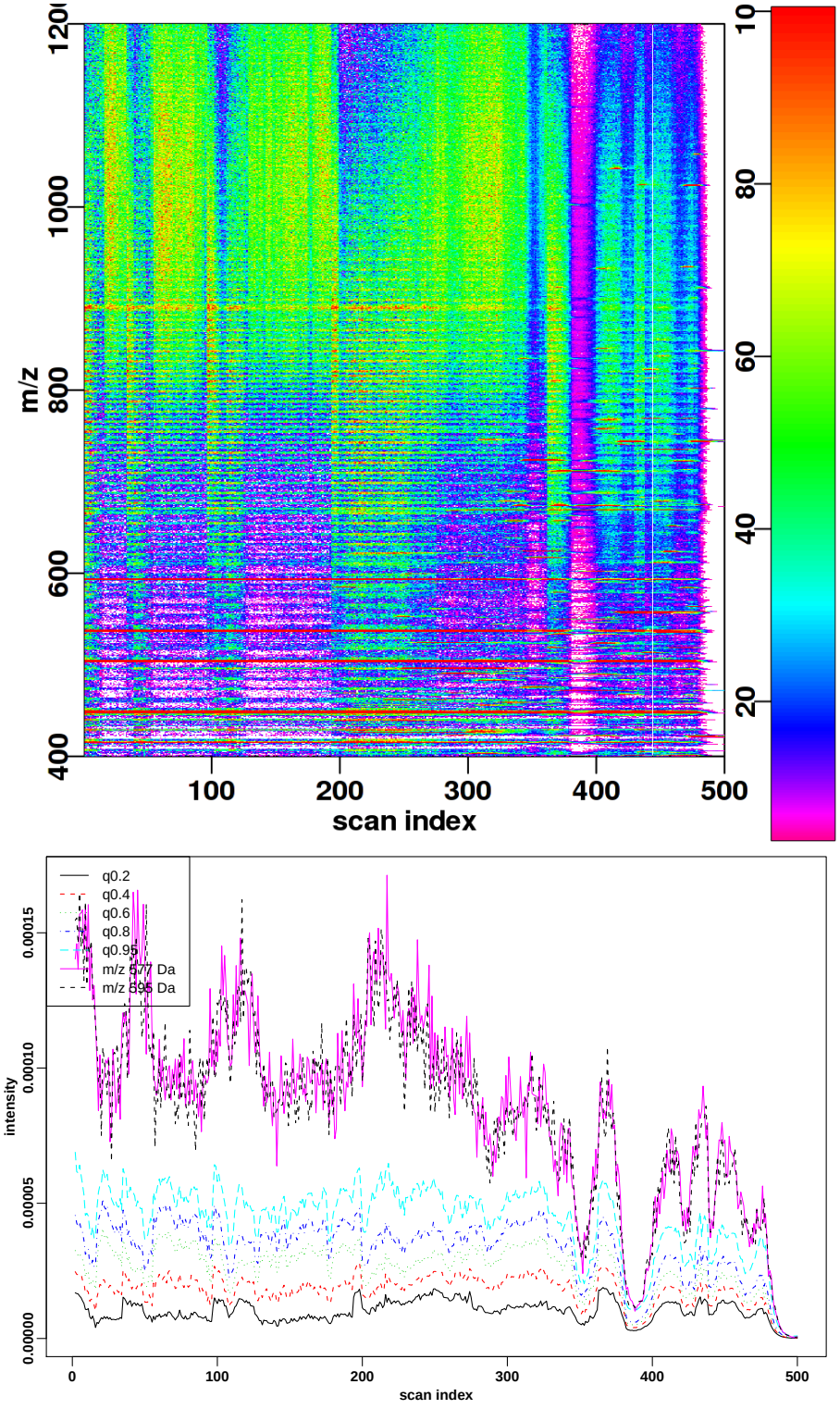

Figure 6.20: After normalization with the euclidean norm.

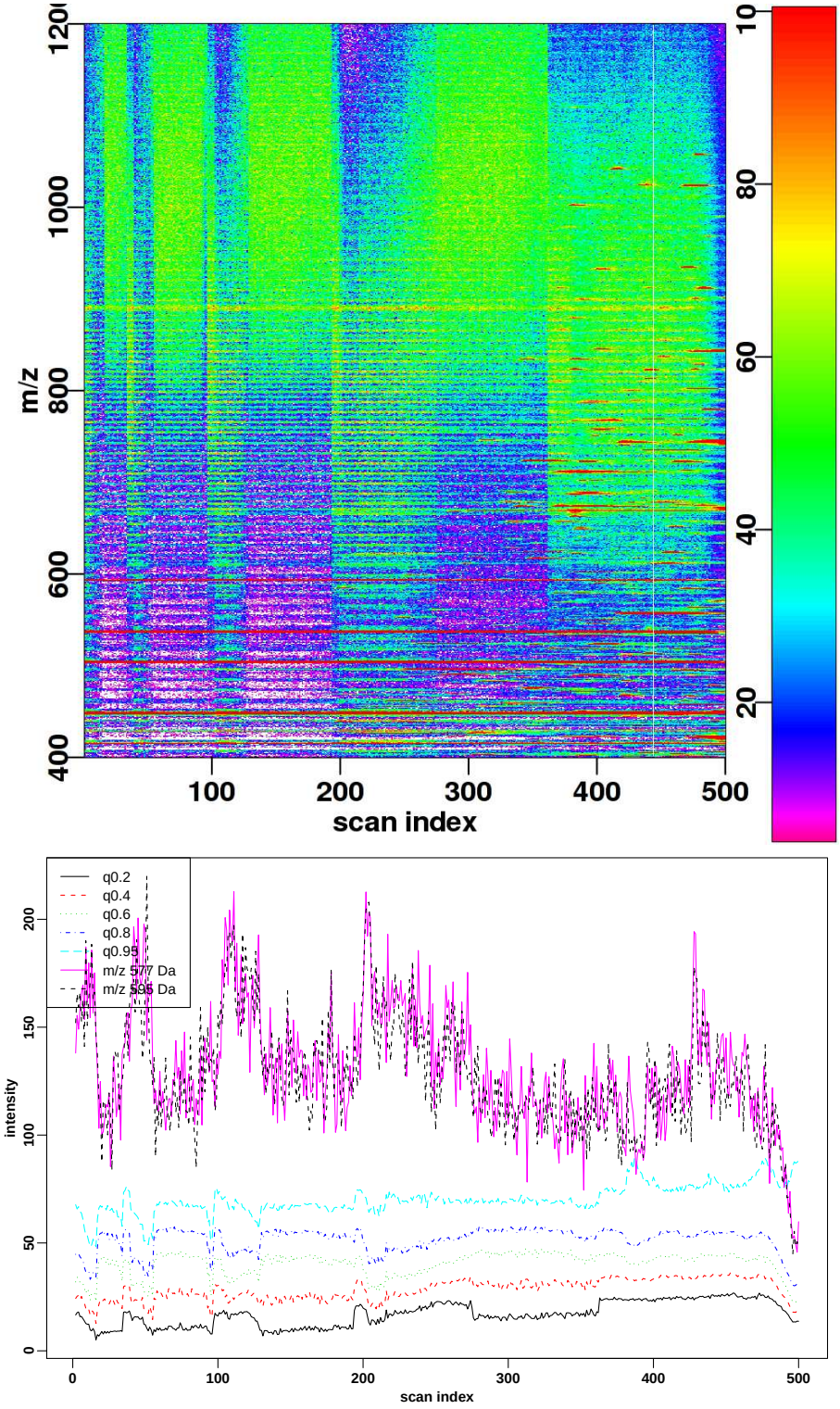

Figure 6.21: After normalization with the Huber m-estimator.

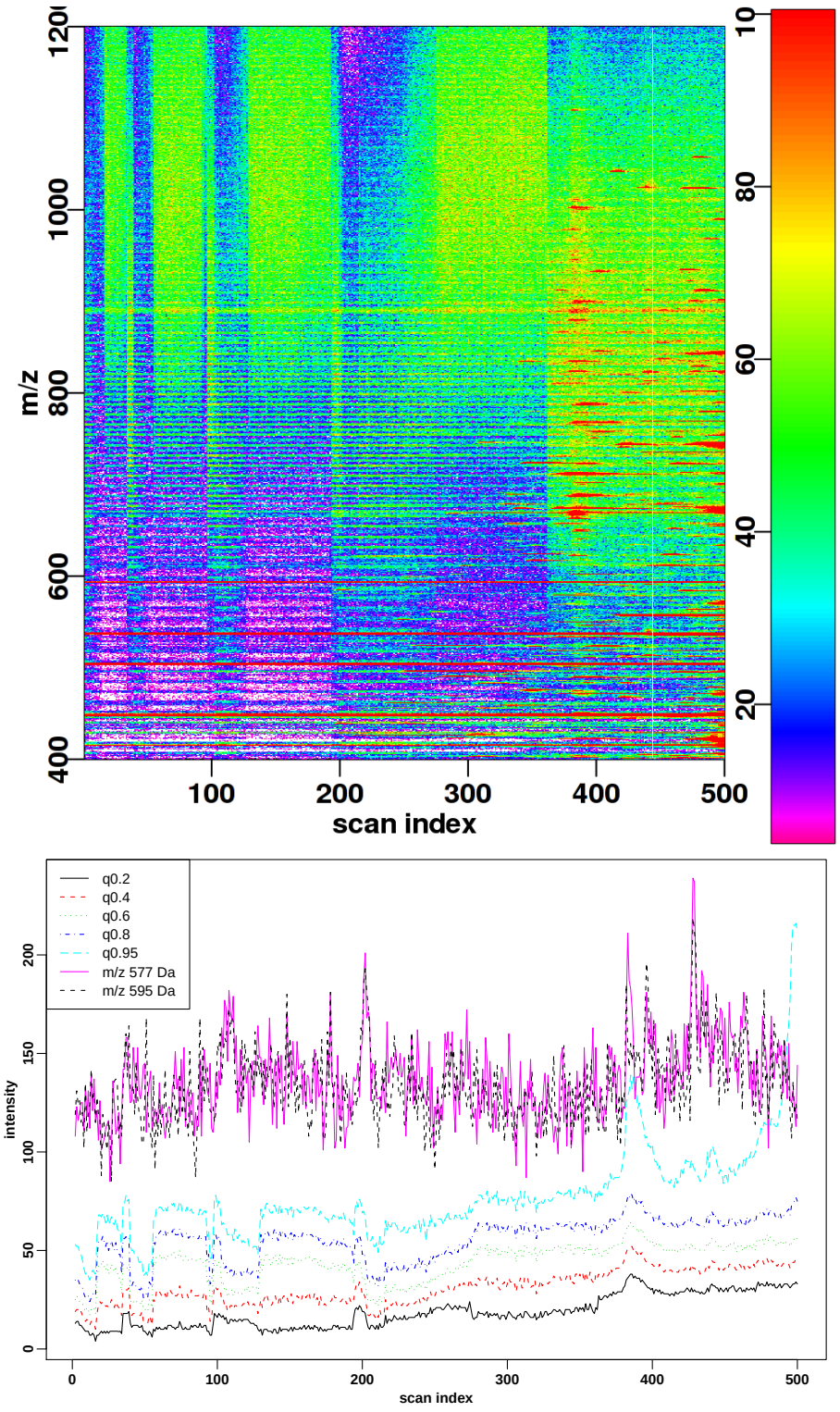

Figure 6.22: After normalization with the median.

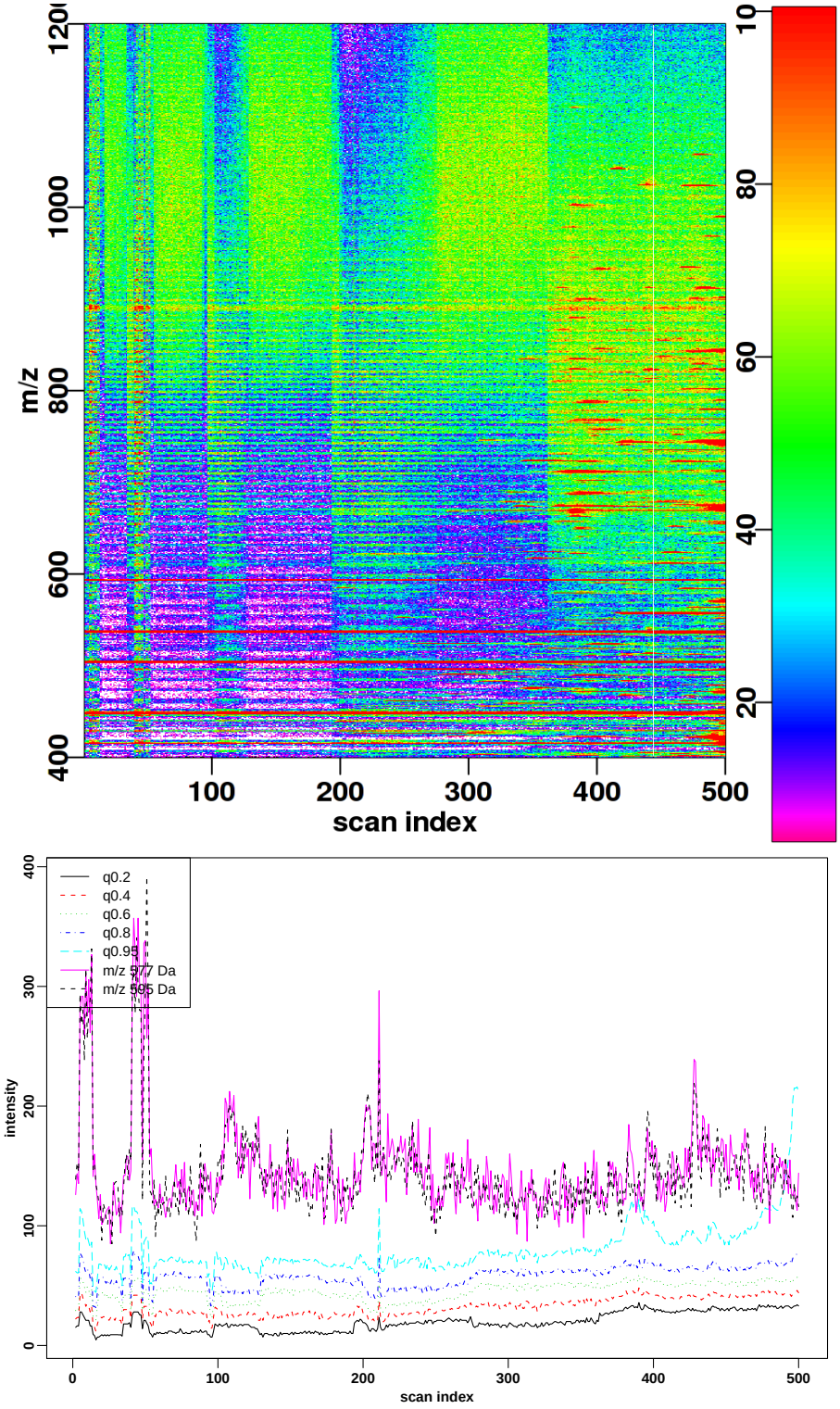

Figure 6.23: After normalization with the weighted mean of the quantiles.

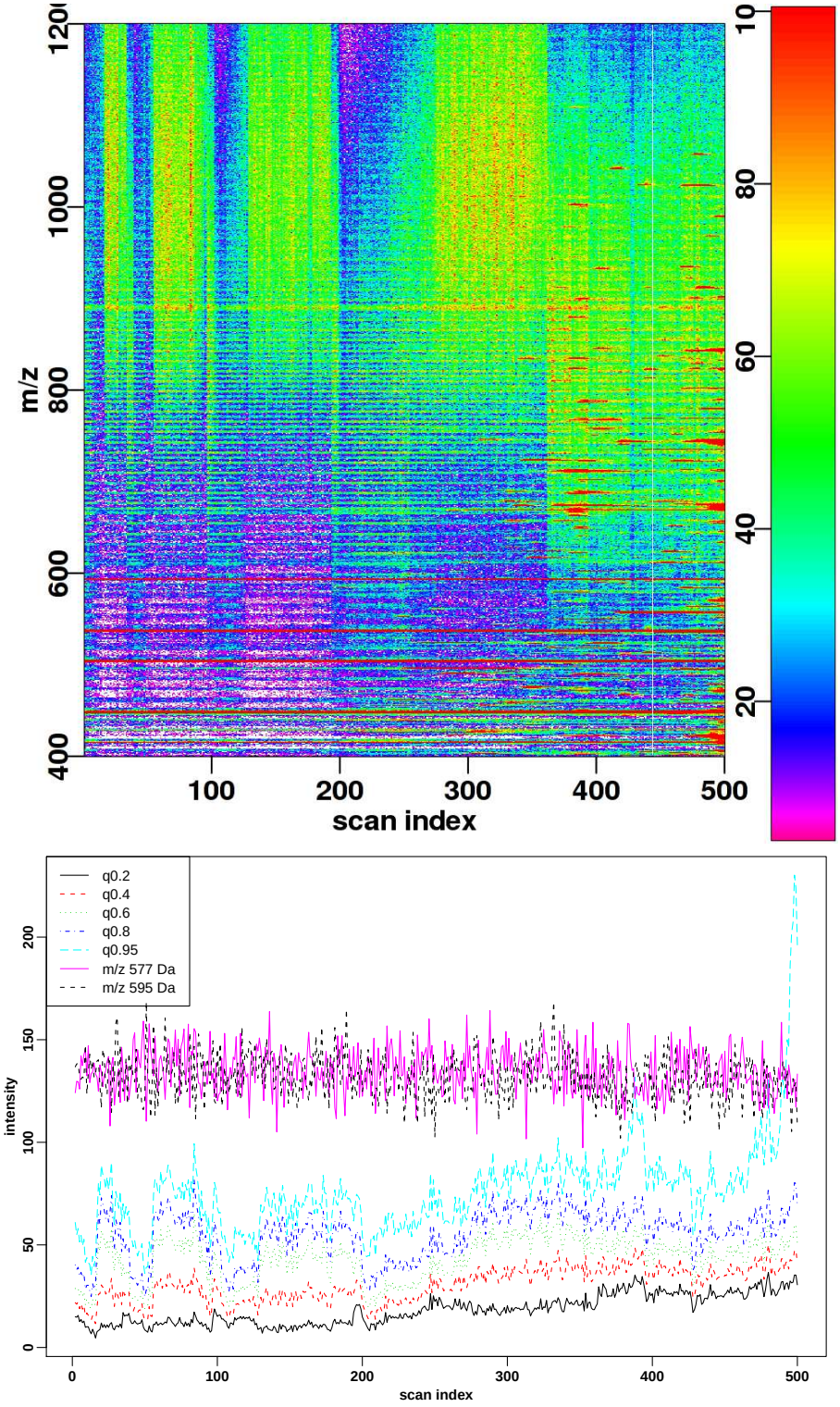

Figure 6.24: After normalization with the weighted mean of contaminant intensity.

6.8.6 Spectrum normalization based on rank one matrices

Problem statement For each mass spectrum with index t, we compute a normalization factor NF_t. We are also interested in estimating the baseline as a function of the m/z ratio m. This corresponds to Problem 3 in the classification.

$$\mathcal{I}_t(m) = \mathrm{NF}_t \left(\mathcal{N}(m) + \mathcal{S}(m)\right)$$

As in normalization based on isolated mass spectra (Section 6.8.5), we consider that the peptide signals are outliers in the background noise distribution. The model becomes:

$$\mathcal{I}_t(m) = \mathrm{NF}_t \; \mathcal{N}(m)$$

and its mean intensity is simply a separable function of retention time and m/z ratio:

$$\mathbb{E}\left[\mathcal{I}_t(m)\right] = \mathrm{NF}_t \; B(m)$$

By estimating the mean noise intensity, we obtain estimates of both the normalization factors and the baseline function.

Due to separability, it is natural to estimate the surface $(t, m) \mapsto \mathrm{NF}_t \; B(m)$ with a rank one matrix. Let $A = (a_{ij}) \in \mathbb{C}^{p \times q}$ denote a complex valued matrix with p rows and q columns. The Froebenius norm of A is $\|A\|_F = \sqrt{\sum i, j |a_{ij}|^2}$. The Singular Value Decomposition (SVD) is a natural tool for finding low rank approximations to the matrix A because of the following theorem (see [GVL96, FT06] and references therein).

Theorem 6.8.1 (Eckart-Young [EY36]). *Let $\mathbb{C}_k^{p \times q}$ be the set of complex valued matrices of rank lower than k. Let $A_k = \sum_{i=1}^{k} s_i \, u_i \, v_i^*$ where s_i, u_i and v_i are the singular values, left-singular and right-singular vectors in the SVD decomposition of A. The matrices A_k verify*

$$\min_{X \in \mathbb{C}_k^{p \times q}} \|A - X\|_F = \|A - A_k\|_F$$

In particular, A_1 is a rank one approximation of A. It can be computed in polynomial time.

To our knowledge, the theory of rank one approximation of matrices has been focused on special types of matrices such as symmetric Toeplitz matrices and Hankel matrices [CFP03]. Singular Value Decomposition seems to be the only method for approximation of arbitrary matrices with an explicit algorithm for practical computation, as opposed to optimization. Optimization methods were not considered in this study because of the large size of LC/MS images.

Update Methods based on Non-negative Matrix Factorization (NMF, [HVD08]) look promising but were not considered in the course of this thesis.

Results Figure 6.25 shows the LC/MS image along with the estimated rank one model A_1. The model appears to faithfully represent the background noise intensity, and also takes into account both changes in normalization and baseline intensity.

The rank one model has several shortcomings:

- The SVD decomposition is an adjustment based on the Froebenius norm and is sensitive to the presence of outliers, i.e. to the presence of peptide signals.
- Outside iterative optimization, we are not aware of a method to compute a solution to the minimization problem in Theorem 6.8.1 with a different norm or distance than the Froebenius norm.

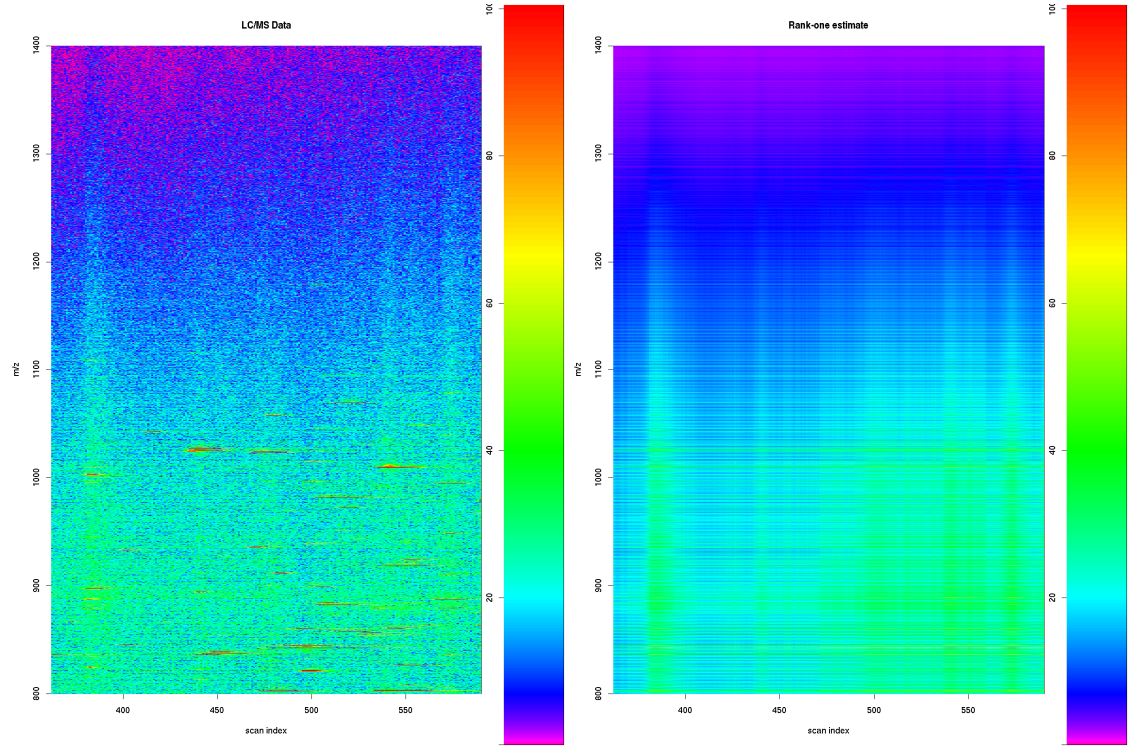

Figure 6.25: Estimation of the normalization factors and the baseline with the Singular Value Decomposition (SVD). The left panel shows the data, and the right panel shows the base surface estimated with the rank one SVD model.

- The rank one model distributes the contribution of peptide signals over the baseline and the normalization factors. There is no way to particularize the baseline coefficients or the normalization factors.
- We expect the SVD model to fit too closely to the LC/MS image, i.e. we anticipate an overfitting problem.

6.8.7 Validation of normalization methods

6.8.7.1 Ideal situation

In the ideal situation, we would have a data set where the normalization factors are known for each mass spectrum. This is not realistic in practice and we discuss the validation criteria used in Section 6.8.5 and the technical difficulties to generate an ideal data set.

Remark In Chapter 7, we build a synthetic data set with known normalization factors to precisely evaluate the performance of the *compatible values estimator* and compare with quantile normalization and TIC normalization (proposed in 6.8.5).

6.8.7.2 Contaminants

The observed contaminants in the chosen data set were identified by our colleagues at the mass spectrometry platform of the Pasteur Institute as molecules liberated by the plastic tubing. The likely culprit is a polymer of dimethylsiloxane (DMS, mass 74 Da) which appears at m/z ratios of the form $1 + 74k$ like 519 Da and 593 Da. This is coherent with molecules of the form $(\text{DMS})_k H^+$.

In section 6.8.5, we used 595 Da, which corresponds to the +2 isotope but appeared more intense in the LC/MS image, and 577 Da which corresponds to a neutral loss of water (-18 Da).

For validation in practice, contaminant traces are reasonable candidates as indicators of the true normalization factor. As the contaminant is thought to be liberated by the plastic tubing or the chromatography column, we made the hypothesis that it is not separated by chromatography and thus liberated at the same rate over time. The figures in Section 6.8.5, and in particular Figure 6.24, are coherent with this hypothesis, but not to the point that it is usable for proper validation.

Instead of relying upon contaminants that are naturally in the LC/MS image, is it possible to induce contamination ?
There are technical difficulties to properly introduce a contaminant in the analysis platform. If it is introduced with the sample before LC/MS, we cannot guarantee that its signal intensity is constant in time because it may interact with the LC column. The contaminant must be introduced before ionization because it must be an indicator of what is going on in the sample.
The only method that we came up with is to perform a post-column infusion of a contaminant. In that case, the contaminant is mixed with the eluent from the liquid chromatography just before electrospray ionization. This mixing step is a challenge in microfluidics [6] because of the small size of the tubes involved. We did not get such a data set.

6.8.7.3 Stationary background distribution

We compared our normalization methods based on the hypothesis that the background noise distribution is the same in each mass spectrum. This is verified to some extent before normalization as can be seen in Figure 6.17. We hoped to improve the situation with our methods, even if Figure 6.16 shows that the background noise patterns may be very different in successive mass spectra.

The data set in [KEH+08] that is used in Chapter 8 is much better behaved, and the stationary hypothesis is more reasonable in that case. However, it does not contain visible contaminants, so validation with both the quantiles and the contaminants was not possible.

The promising results obtained with the rank one / SVD model proposed in Section 6.8.6 also suggest that the background noise distribution is independent of retention time, and that the variations observed are only related to the normalization factors. Consequently, after normalization, we may assume that background noise is identically distributed in a horizontal line. This hypothesis will be used in Chapter 8 for detecting protein signals.

6.8.8 Summary

We propose spectrum normalization in addition to global normalization as a means to remove experimental variations in the multiplicative factor between protein concentration and measured intensity. Several experimental reasons motivate the use of spectrum-specific normalization factors such as competitive ionization and automatic gain control. These variations can be observed in LC/MS images as vertical stripes.

As only global normalization is considered in the literature, we provide a classification of the problems in spectrum normalization. We have studied in particular Problem 1 where normalization is performed in ideal conditions (independence and stationarity), and Problem 3 where the background noise is not uniform in a mass spectrum.

[6] Microfluidics deals with the behavior, precise control and manipulation of fluids that are geometrically constrained to a small, typically sub-millimeter, scale.

Two families of methods are considered. In Section 6.8.5, we describe methods that only consider isolated mass spectra and try to match the background noise distribution to a pre-specified level. In Section 6.8.6, we describe methods based on rank one approximation of matrices that try to model the base surface and provide the spectrum specific normalization factors and the baseline at the same time.

Even though the results look promising, we have been unable to properly validate the proposed methods due to the lack of a suitable validation data set as discussed in Section 6.8.7. As a consequence, we have sought for a noise model with explicit normalization factors. The results of this model-based approach are described in Chapter 7.

6.9 Conclusion

An LC/MS image represents ion intensity as a function of retention time and m/z ratio. These are very large images, on the order of a gigabyte each, which require many steps of preprocessing before identification and quantification of the peptide signals within.

All the components of LC/MS images suffer from systematic sources of variation related to technical difficulties. Retention times are difficult to reproduce because they are sensitive to experimental conditions. M/z values need careful calibration procedures to improve accuracy. Intensity values measured on LC/MS platforms are contaminated with chemical noise and other sources of variation that affect protein quantification.

In this chapter, we presented the types of data recorded on LC/MS platforms, and the types of signals, noise and contaminant signals to expect in LC/MS images. We have presented algorithms to correct retention time, m/z ratio and intensity. In particular, Section 6.4 is extracted from our work published in [VLTTK$^+$08], and Section 6.8 contains unpublished original work. In particular, variations in the gain parameter and the ionization efficiency corrected in Section 6.8 can remove biases for alignments methods based on profile data (described in Section 6.4.1.1).

The experiments in Section 6.8 suggest two types of developments. First, for spectrum normalization, the proposed methods are not able to estimate the normalization factor and the mean intensity of the background noise separately, but only their product. The model-based approach presented in Chapter 7 was developed to address this problem.

Second, after normalization, the background noise process becomes stationary in the retention time direction. Therefore, it is easy to evaluate the errors in feature detection based on this hypothesis. This is the basis for the feature detection methods proposed in Chapter 8.

Chapter 7

A noise model for LC/MS data

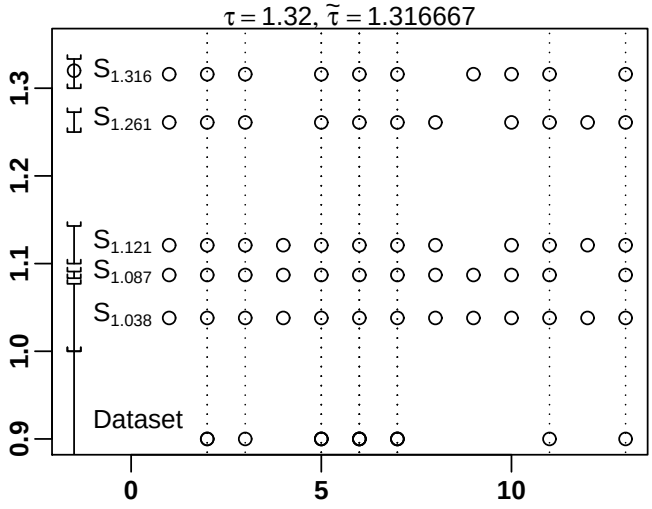

Figure 7.1: Comparison of a few compatible lattices and the dataset.

Models for the background noise in LC/MS data are scarce. Preliminary studies have shown that random noise behaves like Poisson shot noise [ARL+04]. While there are initiatives to model the noise behavior as a function of mass-to-charge ratio [HNR02, MCA+05], these are mainly models for the baseline component. [SKBM04, SMKM07] propose to model the background noise with ARMA processes and estimate the power spectral density. To our knowledge, no model has been proposed from first principles, i.e. based on the knowledge of the physical processes underlying the experiment.

In this chapter, we present a semi-parametric noise model for data acquired on a TOF mass spectrometer. In particular, the normalization factors studied in Section 6.8 are explicitly modeled and we propose several methods to estimate them.

The proposed model is a first attempt at defining a model from first principles and still needs to be extended and validated further. The contents of the chapter are extracted from a manuscript in preparation [LTT08].

7.1 Introduction

7.1.1 Ion Detectors

We consider detectors similar to microchannel plate detectors that are used in most mass spectrometers [Wiz79] (see Chapter 3 for details on ion detection). When an ion hits the detector plate, it produces an analog signal that is amplified, quantized, then reported to the computer. The level of quantization is quite high as there may be only $2^{11} = 2048$ levels[1] in some instruments [SWB03], and small signals as well as chemical noise are strongly affected by quantization effects.

Specific difficulties have appeared with high-throughput analyses of biological material. In particular, biological samples may contain trace amounts of molecules of interest. These are difficult to distinguish from chemical noise which produces patterns similar to real signals [ARC+03, KC02, WPP96]. In 2004, [ARL+04] suggested Poisson-like behavior for the ion intensity based on a linear relationship between the mean and variance of the noise. This linear relationship suggests that the amplification factor of the detector may be unaccounted for in the data set.

Photomultipliers Photon multipliers are devices for counting photons instead of ions. When a photon hits the scintillator, the plate liberates an electron, that electron hits a dynode and liberates additional electrons. Several stages allow for large amplification of the light signal. Photon multipliers are not usually employed in mass spectrometers. More information on these devices can be found in technical reports [2] as well as in [Tan82, MKG03, Ind89].

7.1.2 Summary of the mathematical results

We start from a general model for the intensity $\mathcal{I}(t, m)$ at retention time t and m/z ratio m:

$$\mathcal{I}(t, m) = \tau(t, m) \left(N(t, m) + \textsc{Pep}(t, m)\right)$$

where $\tau(t, m)$ is the gain of the ion detector, $N(t, m)$ is the random number of incident ions and $\textsc{Pep}(t, m)$ is the peptide signal.

In Section 7.2, we simplify the previous model into $X = \lfloor \tau N \rfloor$ where X is a model for the intensity in a mass spectrum and $\lfloor\rfloor$ denotes the floor function. In the following, we will consider each mass spectrum in the LC/MS image independently from the others, and estimate τ and the distribution of N inside a mass spectrum.

In Section 7.3, we discuss the different mathematical situations that arise in the proposed model. We explain why estimation is trivial with negligible quantization error, and why it is impossible with $\tau \leq 1$. The rest of the chapter focuses on the distinguishable case, i.e. $\tau > 1$.

In the distinguishable case, we first study upper bounds for τ in Section 7.4. We then define the notion of *compatible values*, and propose an optimal estimator for τ in Section 7.5.

The performance of the compatible values estimator is compared to three other estimators in Section 7.6. These estimators provide poor results compared to the optimal estimator, but provide insight on the practical issues and the mathematical problem.

Section 7.7 proposes additional mathematical results, and an application of the compatible values estimator to spectrum normalization.

[1] Single precision floating point numbers have $2^{24} \sim 16.10^6$ levels of precision plus sign and exponent.
[2] `http://www.electrontubes.com/technical-information/`

7.2 Modelisation

7.2.1 Practical difficulties

TOF instruments measure the signal intensity on an integer scale, as described in Chapter 6. Moreover, the integer values are often left-censored, i.e. values below 3 are filtered out of the data set. These characteristics have unexpected consequences on signal processing and statistical procedures:

- The mean intensity of the noise is tricky to estimate because of the peptide signals (outliers) and the removal of low-intensity values.
- Quantization errors are expected to have a major contribution to any estimator, e.g. quantiles are necessarily integers.
- The median intensity of most mass spectra is zero, probably because of the truncation of low-intensity values.
- The distribution of chemical noise and that of signal intensity have a significant overlap because of low-intensity peptide signals.

We do not expect a very accurate noise model in this context.

In spite of all these difficulties, we have studied the possibility to estimate the gain parameter of the ion detector and the noise distribution. The results presented in this chapter are uncommon, and fall at the interface between parametric estimation in statistics and discrete mathematics.

As explained in Section 7.2.2, the parameters of the model are estimated inside a mass spectrum, and independently from the other mass spectra. This is because the model is intended to study the variations in the normalization factor and the background noise distribution over time. As such, it may be used to validate the hypotheses in Section 6.8.

7.2.2 Mathematical model

We propose a semi-parametric model for the intensity $\mathcal{I}(t, m)$ as a function of retention time t and m/z ratio m in LC/MS images. Let $\tau(t, m)$ denote the normalization factor, $N(t, m)$ denote the background noise and $\textsc{Pep}(t, m)$ denote the peptide signal. The recorded intensity is thus equal to

$$\mathcal{I}(t, m) = \tau(t, m)\,(N(t, m) + \textsc{Pep}(t, m))$$

Considering the peptide signals as outliers in the noise distribution, we neglect them [3] and obtain:

$$\mathcal{I}(t, m) = \tau(t, m)N(t, m)$$

In order to study the variations of τ and N at different retention times, we will estimate them with a fixed t, that is to say inside a mass spectrum. We make the assumption that τ does not depend on m because it models the global normalization factor in a mass spectrum. In addition, we make the simplifying assumption that N is independent of the m/z ratio m. In conclusion, we are in the hypotheses of normalization problem 2 (see Section 6.8.2 page 113).

$$\mathcal{I} = \tau N$$

Considering the practical difficulties in Section 7.2.1, we introduce a quantization error term [4] with the floor function $\lfloor\rfloor$. As the ion detector is capable of counting single ion events, we model N as an integer-valued random variable. We obtain the following observation model:

$$X = \lfloor \tau N \rfloor$$

[3] Peptide signals are not completely neglected. They are simply not modeled. We will use robust methods to deal with the deviations from the proposed model caused by peptide signals.

[4] Similar results can be obtained with other quantization operators such as rounding (not shown here).

Recall that $\lfloor x \rfloor$ is defined as the unique integer in $\mathbb{Z}$ such that $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$. Equivalently, $\lfloor x \rfloor$ is the largest integer that is lower or equal to x.

The observation model is associated with the statistical structure $(\mathbb{N}, \mathfrak{B}, \mathbb{P}_\theta)$, with $\theta = (\tau, N)$ where τ is a positive real number and N is a probability distribution on $\mathbb{N}$. The statistical problem is to estimate the real-valued parameter τ and the distribution of N, so it is a mixture of parametric estimation and non-parametric estimation. The statistical model is not regular, therefore many tools from classical statistics are not applicable such as Fisher information (see [Mil01] for details).

7.2.3 A priori distribution of N

The results in Section 7.5 show that the estimation of τ and N can be decoupled completely. The only required assumption on N is that it is integer-valued. In that situation, N can be any type of process, independent or not, stationary or not, and can include the peptide signals. The estimation procedure is also robust to left-censored data.

In this section, we briefly discuss our a priori expectations on the distribution of N for several reasons:

- to help the reader understand the components of the model,
- to justify the usage of a Poisson distribution in the simulations,
- to show the relationship with the overdispersion problem and the double Poisson family,
- and also to suggest possible estimation procedures for N, which is not tackled in this thesis.

We believe a priori that N is Poisson distributed as suggested in [ARL$^+$04] because it models rare events (ion counts). A simple estimator for τ and the Poisson parameter λ is

$$\begin{cases} \hat{\tau} = \dfrac{\mathrm{Var}(X)}{\mathbb{E}[X]} \\ \hat{\lambda} = \dfrac{\mathbb{E}[X]}{\hat{\tau}} \end{cases}$$

This estimator deduces the estimated parameters $(\hat{\tau}, \hat{\lambda})$ from the moments of X. This estimator is not very precise. In practice it is not usable because of the peptide signals, which affect the estimation of $\mathrm{Var}(X)$.

The normalization factor τ can be interpreted as an overdispersion parameter affecting Poisson distributed observations. This has been studied extensively in the framework of generalized linear models [MN89, NW72] and common methods include negative binomial regression [Hil07, Law87] and exponential dispersion models [Efr86, Fit97, GD90].

In the model under consideration, a generalized Poisson approach seems better suited than negative binomial regression because the variance of X depends on the mean via $V(X) = \tau\mathbb{E}[X]$. In Section 7.6.2, we apply double exponential families as presented by [Efr86]; these were used by [ABF99] for regression in a similar dataset with the model $X'_i = \tau N_i$ where N_i is Poisson distributed with varying parameter μ_i. However, quantization effects are not taken into account in generalized linear models, and poor parameter estimates are obtained by this approach. Moreover, we wish to confirm the Poisson hypothesis by using non parametric estimation.

7.3 Mathematical situation and types of problems

7.3.1 Trivial case

With negligible quantization error, the observation model becomes $X = \tau N$. Estimation of τ is trivial; all that is required is to observe the event $\{N = 1\}$ i.e. $\{X = \tau \times 1\}$, or the two events $\{x_1 = \tau i\}$ and $\{x_2 = \tau(i + 1)\}$ and compute the difference $x_2 - x_1$. To recover N, it then suffices to consider X/τ. The quantization error may be neglected when $\tau \gg 1$ in the observation model $X = \lfloor \tau\, N \rfloor$ and the previous estimates provide τ with a precision on the order of the quantization error.

7.3.2 Distinguishibility

In the general case, we can recover the samples of N from the samples of X when the mapping $x \mapsto \lfloor \tau\, x \rfloor$ is injective. The inverse mapping is $y \mapsto \lfloor y/\tau \rfloor + 1$. We call this situation the *distinguishable* case. It occurs if and only if $\tau \geq 1$ (see proof in the Appendix, Prop 7.9.1). In this situation, the semi-parametric approach can be separated into parametric estimation of the gain parameter τ followed by non parametric estimation of the distribution of N.

Example 7.3.1.

```
> data = floor(1.32 * n)
% Distinguishible case
> n
 [1]  1  2  3  4  5  6  7  8  9 10
> data
 [1]  1  2  3  5  6  7  9 10 11 13
```

When τ is smaller than 1, the truncation error merges adjacent values of N. In the following example, the events $\{N = 3\}$ and $\{N = 4\}$ cannot be distinguished in the data set. This is because the corresponding observation is $\{X = 2\}$ in both cases.

Example 7.3.2.

```
> data = floor(0.68 * n)
% Non distinguishible case
> n
 [1]  1  2  3  4  5  6  7  8  9 10
> data
 [1]  0  1  2  2  3  4  4  5  6  6
```

In the distinguishable case, it is natural to sort and index the observed values in order to determine the mapping $x \mapsto \lfloor \tau\, x \rfloor$. This is not sufficient in practice because of missing values or outliers which can modify the indexes (See Section 7.6.4).

7.3.3 Consequences

The gain parameter and the law of N have separate effects on X. In the distinguishable case, the distribution function of X is a transformation of the distribution function of N by the mapping $x \mapsto \lfloor \tau\, x \rfloor$. The gain parameter and the truncation error only distort the position of each peak, whereas the relative frequencies are unchanged. Consequently, the support $\mathcal{S}$ of the empirical distribution is sufficient information for estimating τ whereas the empirical frequencies are sufficient information for the distribution of N.

In the non distinguishable case, the set of observed integers is always $\mathbb{N}$ for large samples (see Section 7.5). As a consequence, τ cannot be estimated based on that set alone. A semi-parametric approach is not feasible either. For instance, we cannot estimate the mean $\mathbb{E}[N]$ but only $\mathbb{E}[X] = \tau\mathbb{E}[N]$. To separate τ and N, we have to provide prior assumptions on the distribution of N like a Poisson parametric family.

In the following, we study properties of the set $\mathcal{S}$ of observed integers. This set can be constructed from the dataset in $\mathcal{O}(n\log(n))$ time using sorting for example. The algorithmic complexity of the following algorithms is governed by the size of $\mathcal{S}$, and in particular, the maximum integer in $\mathcal{S}$.

We focus on the distinguishable case, and perform parametric estimation of the gain parameter from a random set of integers. As the support of the empirical distribution is a sufficient statistic for τ, we use the statistical structure $\left(\Omega = 2^{\mathbb{N}}, \mathfrak{T}, \mathbb{P}_\tau, \tau \in]1, +\infty[\right)$ where Ω is the power set of $\mathbb{N}$ and $\mathfrak{T}$ is the exhaustive σ-algebra on Ω.
$(\mathbb{P}_\tau, \tau \in]1, +\infty[)$ is a parametric family of distributions on Ω that is implicitly generated in the following way. For a fixed integer n and fixed but unknown integer-valued random variable N, $\mathbb{P}_\tau$ is the distribution of the random variable $\mathcal{S}$ which is the set of observed integers in an independent identically distributed sample $(X_1, \ldots, X_n)$ of $X = \lfloor \tau N \rfloor$.

7.4 Upper Bounds for τ

The results in this section are based on the following idea. Two points in $\mathcal{S}$ are separated by at least $\lfloor \tau \rfloor$. Consequently, when τ is large, then $\mathcal{S}$ is a sparse set, whereas $\mathcal{S}$ is dense when τ is near 1. For instance, there are consecutive points in $\mathcal{S}$ if and only if $\tau \leq 2$ (see Proposition 7.9.2 in the Appendix).

A better estimate can be obtained by combining more than 2 consecutive points. Let $[\![x, y]\!]$ denote the set of integers between x and y. If $[\![x, y]\!]$ is a subset of $\mathcal{S}$, then $\tau < 1 + \dfrac{1}{y - x}$. Consequently, τ can be estimated by $1 + \dfrac{1}{y - x}$ with a precision on the order of the inverse of the length of the interval $\dfrac{1}{y - x}$. However, this estimator is strongly affected by missing values in $\mathcal{S}$.

Instead of considering all the segments in $\mathcal{S}$, we propose to use the overall density of the set, which is easier to compute algorithmically. Let $\hat{x} = \lfloor \tau\ \hat{n} \rfloor$ denote the largest integer in $\mathcal{S}$. Then $\tau < \dfrac{\hat{x} + 1}{\hat{n}}$. When $\hat{n}$ is unknown (because of potential missing values), let n denote the number of non zero observed integers i.e. the number of elements in $\mathcal{S}$. Then

$$\tau < \frac{\hat{x} + 1}{\hat{n}} \leq \frac{\hat{x} + 1}{n}.$$

Consequently, $\dfrac{\hat{x} + 1}{n}$ is an estimate of τ with precision on the order of $1/n$ (without missing values). As it uses the whole data, it is usually more precise than the previous bound. We will use this in the rest of the paper to restrict the search space for τ.

Let us compare the previous bounds on an example. Suppose that $\tau = 1.32$ and that we observe $\mathcal{S} = \{1, 2, 3, 5, 6, 7, 9, 10, 11, 13\}$.
As there are consecutive integers in $\mathcal{S}$ we obtain $\tau < 2$.

Using the interval $[\![5, 7]\!]$, we obtain $\tau < 1 + 1/2$.
Using the interval $[\![9, 11]\!]$, we obtain $\tau < 1 + 1/2$ as well.
The density upper bound is $\tau < 14/10$.

Remark We only provided upper bounds in this section because lower bounds can only be deduced from the integers that cannot be generated in the model. These are difficult to distinguish from missing values, which are integers that can be generated in the model, but do not appear in the set $\mathcal{S}$ of observed integers.

7.5 Compatible Values

The upper bounds that we proposed in the previous section are easy to compute but rather poor because they only take into account the proportion of observed integers. In this section, we describe an algorithm with higher computational load but which can leverage the information in the location of each observed integer in the data set.

7.5.1 Lattices of Integers

In the observation model $X = \lfloor \tau N \rfloor$ where N is integer valued, only specific integers can be generated. Given a strictly positive real number t, let us define the set of possible values for x as the *lattice associated to* t, i.e. the infinite set of integers $\mathcal{S}_t = \{x = \lfloor t\, k \rfloor, k \in \mathbb{N}\}$. The set of observed integers $\mathcal{S}$ is also called the *empirical lattice*.

With infinitely many observations, the parameter τ is completely characterized by the empirical lattice as the following proposition shows. This justifies that $\mathcal{S}$ is sufficient information for estimating τ.

Proposition 7.5.1 (Equivalence between lattices and numbers)**.** *In the distinguishable case, let t_1 and t_2 denote two real numbers such that $t_1 \geq 1$ and $t_2 \geq 1$. Then $\mathcal{S}_{t_1} = \mathcal{S}_{t_2}$ if and only if $t_1 = t_2$.*

Proof. Obviously, if $t_1 = t_2$ then $\mathcal{S}_{t_1} = \mathcal{S}_{t_2}$. Let us prove the converse, i.e. $\mathcal{S}_{t_1} = \mathcal{S}_{t_2}$ implies $t_1 = t_2$ or equivalently if $t_1 \neq t_2$ then $\mathcal{S}_{t_1} \neq \mathcal{S}_{t_2}$. Suppose that $t_1 < t_2$. There exists $n \in \mathbb{N}$ such that $\lfloor t_1 n \rfloor < \lfloor t_2 n \rfloor$. Either $\lfloor t_2 n \rfloor \notin \mathcal{S}_{t_1}$, in which case $\mathcal{S}_{t_1} \neq \mathcal{S}_{t_2}$, or $\lfloor t_2 n \rfloor = \lfloor t_1 n_1 \rfloor$ with $n_1 > n$. In the latter case, distinguishability implies that there are strictly more elements in $\mathcal{S}_{t_1} \cap A$ than in $\mathcal{S}_{t_2} \cap A$ where A denotes the set of integers $A = [\![0, \lfloor t_2 n \rfloor]\!] = [\![0, \lfloor t_1 n_1 \rfloor]\!]$. □

7.5.2 The Set of Compatible Values

Proposition 7.5.1 is not sufficient for estimating τ because in practice we only observe a finite set $\mathcal{S} \subsetneq \mathcal{S}_\tau$. Consequently we define the notion of compatible lattices and equivalently compatible values. For positive real numbers t_1 and t_2, we say that t_1 is *compatible* with t_2 if $\mathcal{S}_{t_2} \subset \mathcal{S}_{t_1}$. Likewise, a positive real number t is compatible with the data if $\mathcal{S} \subset \mathcal{S}_t$. Being compatible with the data set is a necessary condition for a valid estimator of τ.

The set of values that are compatible with the infinite lattice $\mathcal{S}_\tau$ is adequate for estimating τ because of the following proposition.

Proposition 7.5.2. *τ is the largest real number that is compatible with $\mathcal{S}_\tau$.*

Proof. τ is a compatible value, we only have to show that it is the largest.
Let u denote a real number greater than τ, and let α denote a positive real number such that $\tau < \tau + \alpha < u$. We will prove that u is not compatible with τ by constructing an element in $\mathcal{S}_\tau$ that cannot be in $\mathcal{S}_u$.

Let a denote a positive integer such that $a > \frac{1}{\alpha}$ and $n = \lfloor \tau a \rfloor$.
Suppose that $\mathcal{S}_\tau \subset \mathcal{S}_u$ then n belongs to $\mathcal{S}_u$ and there exists a positive integer a' such that $n = \lfloor ua' \rfloor$.
$a' \geq a$ because in the distinguishable case, a and a' correspond to their indices in the sets $\mathcal{S}_\tau$ and $\mathcal{S}_u$ and $\mathcal{S}_\tau \subset \mathcal{S}_u$.
Moreover, as $\tau \leq u$ we have $\lfloor \tau a \rfloor \leq \lfloor ua \rfloor \leq \lfloor ua' \rfloor$. For all three terms to be equal to n in the distinguishable case requires that $a = a'$.
Consequently, both τ and u lie in the interval $[\frac{n}{a}, \frac{n+1}{a}[$. As a result, $|\tau - u| \leq \frac{1}{a}$ which contradicts the hypothesis $a > \frac{1}{\alpha}$. □

The set of real numbers that are compatible with $\mathcal{S}_\tau$ has an intricate structure. It contains the positive real numbers smaller than 1 and the harmonics $\left\{\frac{\tau}{k}, k \in \mathbb{N}^*\right\}$, but these are not the only values. For example, $4/3$ is compatible with 2 because every even integer can be written as $\lfloor k \times 4/3 \rfloor$, $k \in \mathbb{N}$. Indeed, let k be an even integer. Either k is a multiple of 4, in which case $k = 4i = \lfloor \frac{4}{3} \times 3i \rfloor$, or $k = 4i + 2 = \lfloor \frac{4}{3} \times (3i + 2) \rfloor$.

7.5.3 Estimation with a Finite Lattice

In practice, the empirical lattice is finite and can contain missing values and outliers. We say that an integer is *missing* from $\mathcal{S}$ when it is in the theoretical lattice $\mathcal{S}_\tau$, smaller than $\hat{x} = \max \mathcal{S}$, but not in $\mathcal{S}$. The set of compatible values with $\mathcal{S}$ is a finite union of intervals and compatible values are never isolated. As the data contains less information, the true parameter τ is not the largest compatible value, but it is still maximal in the following sense.

Proposition 7.5.3. *The set of compatible values $\mathcal{C}(\mathcal{S})$ contains exactly $]0, 1]$ and intervals of length at least $1/\hat{x}^2$ where $\hat{x} = \max \mathcal{S}$. In particular, if there are no outliers or missing values in $\mathcal{S}$ then τ belongs to the interval $[a, b[$ such that $b = \sup \mathcal{C}(\mathcal{S})$.*

The proof is based on the following two lemmas.

Lemma 7.5.1. *The set of compatible values contains exactly $]0, 1]$ and a finite number of intervals of the form $[a, b[$ of length at least $1/\hat{x}^2$ where $\hat{x} = \max \mathcal{S}$.*

Proof. Let $t > 1$ denote a compatible value. For each observed value $x \in \mathcal{S}$, there exists an integer n such than $x = \lfloor tn \rfloor$. Consequently, t verifies $t \in [\frac{x}{n}, \frac{x+1}{n}[$. The intersection of the constraints $t \in [\frac{x}{n}, \frac{x+1}{n}[$ for all $x \in \mathcal{S}$ is an interval $t \in [\frac{x_1}{n_1}, \frac{x_2}{n_2}[$. All values $t \in [\frac{x_1}{n_1}, \frac{x_2}{n_2}[$ verify all of the constraints and are thus compatible. The length of this interval is $\frac{x_2}{n_2} - \frac{x_1}{n_1}$ which is at least $1/(n_1 n_2)$. In the distinguishable case, $n_1 < \max \mathcal{S}$ and $n_2 < \max \mathcal{S}$, which implies that the length is at least $1/(\max \mathcal{S})^2$. □

Lemma 7.5.2. *Let t be a positive real number that is compatible with the empirical lattice. Then $t < \dfrac{\hat{x} + 1}{n}$.*

Proof. This follows directly from the density upper bound in Section 7.4. See Proposition 7.9.4 in the Appendix. □

Proposition 7.5.3. To complete the proof, it suffices to show that τ belongs to the interval with largest values. There are only finitely many intervals of length at least $1/\hat{x}^2$ in $[0, \frac{\hat{x}+1}{n_{max}}]$, so there exists such an interval $[a, b[$.
Let $\mathcal{N}$ denote the set $\mathcal{N} = \{n | \lfloor n\tau \rfloor \in \mathcal{S}\}$, i.e. the set of values for the noise process N that generates $\mathcal{S}$. τ belongs to a certain interval $[a', b'[$ which is the intersection of the constraints $\tau \in [\frac{x}{n}, \frac{x+1}{n}[$, for all $x = \lfloor n\tau \rfloor$ in $\mathcal{S}$. We show that $b' = b$, i.e. no positive real is both greater than b' and compatible. Let t such that $b' < t$. For all $n \in \mathcal{N}$, $\lfloor n\tau \rfloor \leq \lfloor nt \rfloor$. As $t \notin [a', b'[$, t breaks at least one of the constraints, that is to say, there is an integer x in $\mathcal{S}$ such that $x = \lfloor n\tau \rfloor < \lfloor nt \rfloor$. x is skipped in $\mathcal{S}_t$ and thus t is not compatible. □

The previous proposition suggests that it suffices to find the largest compatible interval to estimate τ, and this is our proposed estimator $\tilde{\tau}$. More precisely, the set $\mathcal{C}(\mathcal{S}) = \cup_{j=1}^{J} [a_j, b_j[$ is a union of J intervals, with (a_j) and (b_j) increasing sequences, then

$$\tilde{\tau} = \frac{a_J + b_J}{2}.$$

We use the following algorithm to compute $\tilde{\tau}$. This also computes the mapping $x = \lfloor \tau n \rfloor \mapsto n$ and the precision.

- compute the set of observed values by sorting the data set and removing multiple occurrences
- compute the upper bound $\tau < B = \dfrac{\hat{x}+1}{\text{card}\mathcal{S}}$
- find an approximation of the largest compatible value t by testing the compatibility of the real numbers $t_k = B - \dfrac{k}{\hat{x}^2}$
- deduce the indexes from t, that is to say for all $x \in \mathcal{S}$, find i such that $x = \lfloor ti \rfloor$
- compute the interval $[a, b[$ as the intersection of the constraints $t \in \left[\dfrac{x}{i}, \dfrac{x+1}{i}\right[$, for all x in $\mathcal{S}$
- return $\dfrac{a+b}{2}$ as an estimator for τ

7.5.4 Properties of the Estimator

According to the previous results, the estimator performs well when there are no missing values or outliers. Its precision is $(b-a)/2$ and can be computed inside the algorithm. The precision is at least $1/\hat{n}$ where $\hat{n}$ is defined by $\hat{x} = \lfloor \tau \hat{n} \rfloor$, but depending on the value of τ it can reach a precision on the order of $1/\hat{n}^2$. In all cases, the precision is better than the density bound, and there is a lower bound.

If there are missing values or outliers, the algorithm may find an interval of compatible values that does not contain τ. For example, if the dataset is $\{0, 2, 4, 6, 8\}$, a reasonable estimator would answer 2 and not $\tau = 4/3$ with missing values 1 and 5. In practice, such cases are rare, and are related to arithmetic properties of the set $\mathcal{S}$. However, the largest compatible value is never an erroneous answer to the problem. It is a parsimonious answer in the sense that it is the smallest lattice which may explain the dataset.

The estimator is optimal in the sense that the algorithm finds an interval of positive real numbers that are all equally plausible. Given a dataset $(x_1 = \lfloor \tau\ i_1 \rfloor, \ldots, x_n = \lfloor \tau\ i_n \rfloor)$ of size n, there is an interval of compatible values that can generate $(x_1, \ldots, x_n)$ from the same realization $(i_1, \ldots, i_n)$ of N. Let $[a_J, b_J[$ with $b_J = \sup \mathcal{C}(\mathcal{S})$, the following proposition holds.

Proposition 7.5.4. *Given a realization $(i_1, \ldots, i_n)$ of N, all values in $[a_J, b_J[$ generate the same data set $(x_1, \ldots, x_n)$, i.e.*

$$\forall t \in [a_J, b_J[,\ \forall j \in [\![1, n]\!],\ x_j = \lfloor a_J\ i_j \rfloor = \lfloor t\ i_j \rfloor$$

Proof. As in the proof of Proposition 7.5.3, $[a_J, b_J[$ is the intersection of the constraints $x_j = \lfloor t\ i_j \rfloor$. □

The data set does not contain enough information to distinguish the values in $[a_J, b_J[$. In particular, even if the realization $(i_1, \ldots, i_n)$ is given, then the values are not distinguishable. Note that if x_0 is known not to be in $\mathcal{S}_\tau$, then for all integers i, $\tau \geq \frac{x_0+1}{i}$ or $\tau < \frac{x_0}{i}$. These inequalities are not informative because they are already contained in $x = \lfloor ti \rfloor, \forall x \in \mathcal{S}$.

The program is quite fast. First because is relies only on the set $\mathcal{S}$ which is much smaller than the dataset when τ is near 1 and N is independent identically distributed, because repeats of N are discarded. As the following proposition shows, with few missing values, the density bound is precise and the algorithm is quicker.

Proposition 7.5.5. *Let $\hat{x} = \max \mathcal{S} = \lfloor \tau \hat{n} \rfloor$. If there are no missing values, the largest compatible value is found after at most $\dfrac{\hat{x}^2}{\hat{n}} \simeq \tau \hat{x}$ steps. With a small number of missing values $k \ll \hat{n}$, the number of steps is on the order of $\tau^2 \hat{x} \left(k + \dfrac{1}{\tau} \right)$ where $k = \hat{n} - \mathrm{card}\mathcal{S}$ is the number of missing values.*

Proof. The procedure begins at $B = \frac{\hat{x}+1}{n}$, ends before $\frac{\hat{x}}{\hat{n}}$ because $\tau \geq a \geq \frac{\hat{x}}{\hat{n}}$, and proceeds in steps of length $1/\hat{x}^2$. Consequently, there are at most $C = \hat{x}^2(\frac{\hat{x}+1}{n} - \frac{\hat{x}}{\hat{n}})$ steps. Let $k = \hat{n} - \mathrm{card}\mathcal{S}$ denote the number of missing values. We make the following three approximations: $k \ll \hat{n}$, $1 \ll \hat{x}$ and $\tau \simeq \frac{\hat{x}}{\hat{n}}$. Then $C = \hat{x}^2(\hat{x}+1) \left(\frac{k}{\hat{n}\mathrm{card}\mathcal{S}} + \frac{1}{\hat{n}(\hat{x}+1)} \right)$ which can be approximated by $C \simeq \tau^2 \hat{x}(k + \frac{1}{\tau})$. □

Testing for the compatibility of a real t is linear in the size of $\mathcal{S}$, so the whole procedure is at most quadratic. The full set of compatible values can be obtained in cubic time.

7.6 Results and Discussion

7.6.1 Compatible Values Estimator

Figure 7.2 illustrates the compatible values estimator on a simulated dataset. The dataset $\{6, 6, 11, 5, 3, 5, 2, 6, 5, 13, 2, 7, 7, 7, 6\}$ is obtained from the observation model $X = \lfloor 1.32 * N \rfloor$ where N is distributed according to a Poisson random variable with mean 5.5. It is first reduced to the lattice $\mathcal{S} = \{2, 3, 5, 6, 7, 11, 13\}$ and is shown at the bottom.

The vertical axis represents values of τ. The set of compatible values is composed of several intervals and represented on the left. For each interval, we select one compatible value t and represent the lattice $\mathcal{S}_t$. All reals in the same interval generate the same lattice, up to $\max(\mathcal{S})$.

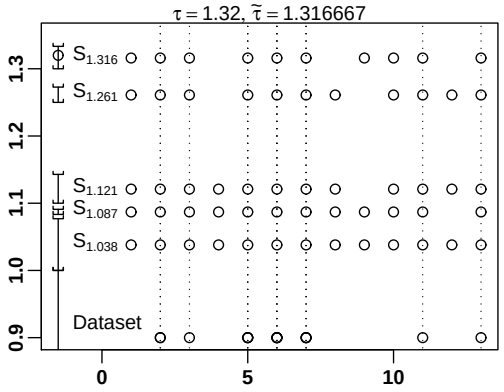

Figure 7.2: Comparison of a few compatible lattices and the dataset.

For comparison, Figure 7.3 displays $\mathcal{S}_t$ for several values that are not compatible with the data. For example, 5 and 11 are in the dataset but not in $\mathcal{S}_{1.2}$.

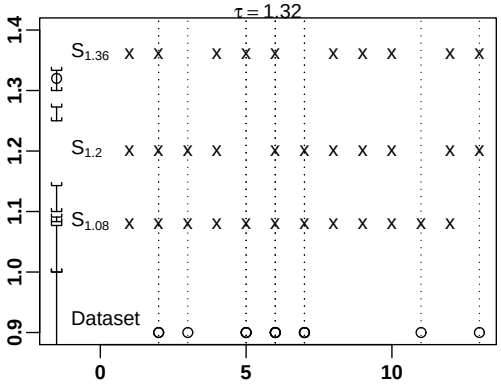

Figure 7.3: Comparison of the dataset and a few lattices that are not compatible.

Two sources of variation affect the estimate $\tilde{\tau}$. First, the estimator is not perfect because the dataset is finite. Second, the data set $\mathcal{S}$ is random. Figure 7.4 shows the performance of the estimator with a fixed dataset ($N \in [\![1, 10]\!]$) for several values of τ. The intervals shown correspond to the intervals of compatible values that contain the largest compatible value.

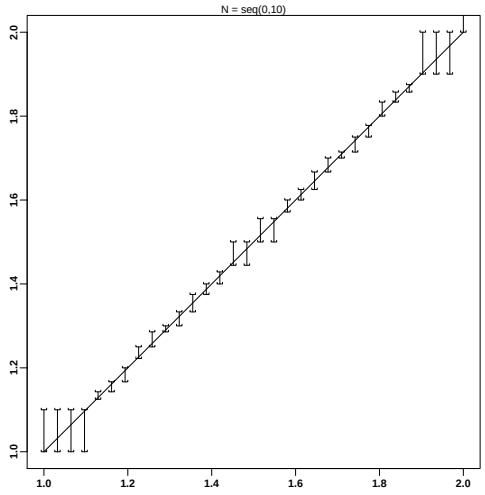

Figure 7.4: Length of the maximal interval for several values of τ with $N \in [\![1, 10]\!]$.

We can see that the precision of the estimator varies with τ. Only the range $[1, 2]$ is shown because the precision only depends on the rest $\tau - \lfloor \tau \rfloor$ modulo 1. Consequently, the absolute precision is roughly constant, whereas the relative precision is $O(1/\tau)$. With small quantization error ($\tau \gg 1$) the estimation problem is easier.

Figure 7.5 shows the distribution of $\tilde{\tau}$ when the dataset $\mathcal{S}$ is the result of 15 samples of $X = \lfloor \tau N \rfloor$ where $\tau = 1.32$ and N is distributed according to a Poisson random variable with mean 5.5. The distribution of the estimator value is obtained from the 200 repeats shown in the bottom of the plot thanks to a kernel estimate, even if the distribution is a sum of Dirac point masses. The drawn interval shows the interval obtained with the (complete) dataset $\{\lfloor 1.32 * n \rfloor, n \in [\![1, 13]\!]\}$.

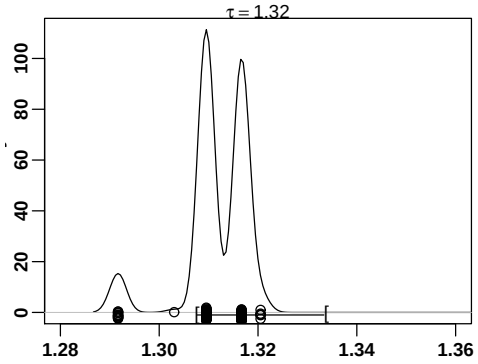

Figure 7.5: Kernel density estimate of the distribution of the compatible values estimator on a random dataset.

7.6.2 The Double Poisson Family

According to the following simulation study, maximum likelihood estimation in the framework of double exponential families does not lead to an estimate of the overdispersion parameter τ with sufficient precision. Let us first briefly recall the results from [Efr86]. Let $g_\mu(y) = e^{-\mu}\mu^y/y!$ denote the distribution function of a Poisson random variable with mean μ. The double Poisson distribution with parameters θ, μ is defined as:

$$\begin{aligned} f_{\theta,\mu}(y) &= c(\theta,\mu)\theta^{1/2}\left\{g_\mu(y)\right\}^\theta\left\{g_y(y)\right\}^{1-\theta} \\ &= c(\theta,\mu)\left(\theta^{1/2}e^{-\theta\mu}\right)\left(\frac{e^{-y}y^y}{y!}\right)\left(\frac{e\mu}{y}\right)^{\theta y} \end{aligned}$$

where c is a normalization constant.

Maximum Likelihood Estimation leads to the following estimators. Let $(Y_1, \ldots, Y_n)$ be independent identically distributed random variables with distribution $f_{\theta,\mu}$, then

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^n Y_i \\ \hat{\theta} &= \frac{n}{2\sum_{i=1}^n I(Y_j,\mu)} \end{aligned}$$

where $I(\mu_1,\mu_2) = \mu_1(\log(\mu_1) - \log(\mu_2)) - (\mu_1 - \mu_2)$.

Let $Y_{\theta,\mu}$ be a random variable with distribution function $f_{\theta,\mu}$. According to [Efr86], $Y_{\theta,\mu}$ has approximately the same distribution as X/θ where X is Poisson distributed with mean $\mu\theta$. With Poisson distribution for the ion counts, our observation model becomes $Y = \lfloor \tau N \rfloor$ where N is Poisson distributed with mean λ. Consequently, estimates for τ and λ can be deduced from $\hat{\theta}$ and $\hat{\mu}$ with the following relations:

$$\begin{aligned} \hat{\tau} &= \frac{1}{\hat{\theta}} \\ \hat{\lambda} &= \hat{\mu}\hat{\theta} \end{aligned}$$

The double Poisson distribution is a correct approximation of the distribution of X/θ for large μ, and in that case, $\hat{\tau}$ and $\hat{\lambda}$ are unbiased estimates of τ and λ. The standard deviation of $\hat{\tau}$ is

$\frac{\tau\sqrt{2}}{\sqrt{n}}$. Figure 7.6 shows the distribution of $\hat{\tau}$ with flooring noise, i.e. in the model $Y = \lfloor \tau N \rfloor$ (solid line) and without flooring noise in the model $Y = \tau N$ (dotted line). The plot was generated with 2000 repeats with data sets of size 500. We observe a large standard deviation compared with the compatible values estimator, even on a much larger data set. With large values of μ, truncation has limited effect on the estimate.

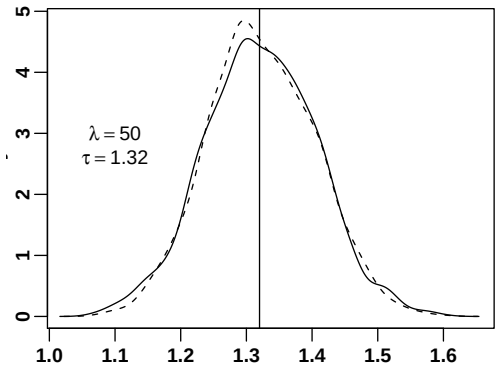

Figure 7.6: Kernel density estimate of the distribution of the double Poisson estimate of τ with flooring noise (solid line) and without (dotted line).

For modeling rare ion count events, we need to study the estimators with small values of λ and τ. In that case, $\hat{\tau}$ is strongly biased for both models as shown on Figure 7.7. This implies that the approximation is not suited to this range of parameters, and that the flooring noise makes a significant difference there. Figure 7.7 was generated using 2000 repeats with data sets of size 500. For comparison, we show the optimal interval obtained by the compatible values estimator on the data set $\{\lfloor 1.32 * n \rfloor, n \in [\![1, 13]\!]\}$.

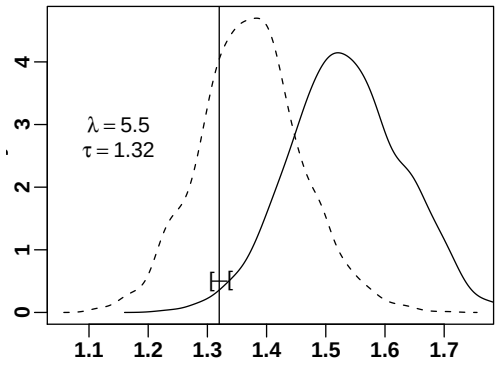

Figure 7.7: Kernel density estimate of the distribution of the double Poisson estimate of τ with flooring noise (solid line) and without (dotted line).

7.6.3 Fourier Estimator

The parametric estimation problem can be solved using Fourier transform thanks to the following remark. From the set $\tau\mathbb{N}$ we construct the signal $f : t \mapsto \sum_{k\in\mathbb{N}} \delta(x - \tau k)$ where δ denotes the Dirac function, that is to say a periodic series of pulses. Likewise, we define the estimator $1/\tau_F$ as

the maximum of the Fourier transform of the quasi-periodic signal $f : t \mapsto \sum_{\lfloor \tau k \rfloor \in \mathcal{S}} \delta\left(x - \lfloor \tau k \rfloor\right)$.

As τ can be seen as a quasi-period, our estimation problem is closely linked to the "harmonic retrieval problem". Many approaches have been proposed in that domain and the main focus is on the estimation of the Power Spectral Density [Hay96]. However, the signal is usually perturbed by additive noise whereas in this paper we consider a distortion of the time axis.

We use the following algorithm:

- sample the signal f at the points $x_i = i$ for the integers i in $[\![0, \max(\mathcal{S})]\!]$
- compute the Discrete Fourier Transform
- compute an upper bound using Proposition 7.9.4 : $\tau < B = \dfrac{\hat{x}+1}{n}$
- find the frequency with highest absolute Fourier coefficient
- return the corresponding period (inverse of the frequency)

This estimator has a precision that corresponds to the sampling rate in time space around the true value. In the Fourier space, the sampling rate is uniform with steps of length $1/\max(\mathcal{S})$ which is equivalent to $1/(\tau \times \hat{x})$. In the time space, as $P = 1/f$, then $\Delta P = -\Delta f/f^2$ and the sampling rate is non uniform. For $f = 1/\tau$ we obtain the precision of the Fourier estimator as $\tau/\hat{x}$. This suggests that the precision decreases with τ. However, the signal frequency $1/\tau$ is near $\hat{x}/\max(\mathcal{S})$ which is one of the sampling points. As a result, in practice, the absolute precision is on the order of $1/\hat{x}$ and independent of τ.

Figure 7.8 shows in the frequency and period space the Fourier transform of the quasi-periodic signal obtained from the dataset $N \in [\![1, 10]\!]$. The vertical line corresponds to the upper bound from Proposition 7.9.4.

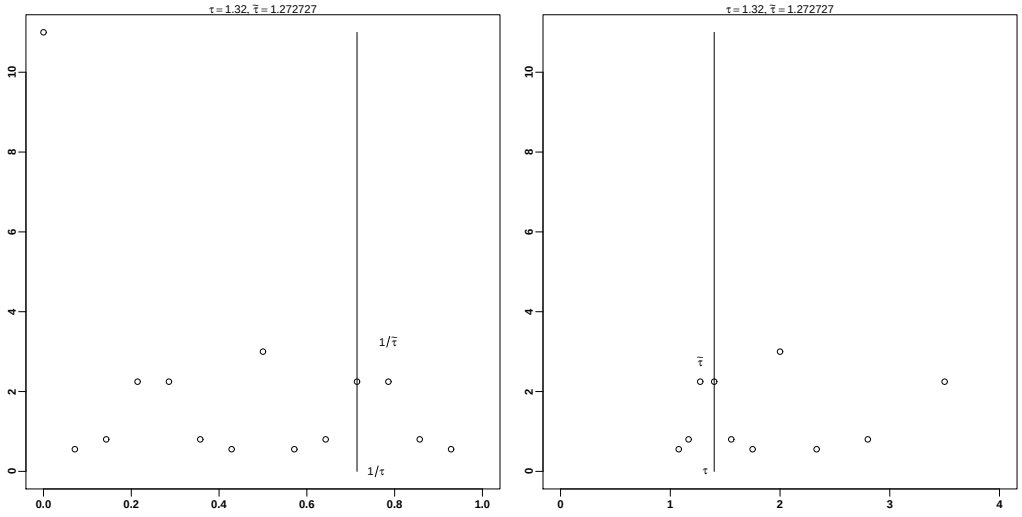

Figure 7.8: Fourier transform of the quasi-periodic signal, in Fourier space (left) and period space (right). The vertical line shows the upper bound from Proposition 7.9.4.

Remark When oversampling by a factor k, i.e. sampling at the points $x_i = \dfrac{i}{k}$ for the integers i in $[\![0, \max(\mathcal{S}) \times k]\!]$, the harmonics of 1 Hz increase in magnitude. Therefore it is necessary to weed out the frequencies above 1 Hz in the distinguishable case. Moreover, oversampling increases

the maximum frequency that can be represented in the Fourier space and does not improve the precision of the estimator.

On a random dataset, the Fourier estimator suffers greatly from missing values. Figure 7.9 shows the distribution of τ_F with 200 simulations and a dataset of size 15 where N is distributed according to a Poisson random variable with mean 5.5. The precision of the estimator is much worse than the compatible values estimator (see the plotted interval). The Fourier estimate τ_F is compatible with the dataset in only about 1% of the simulations.

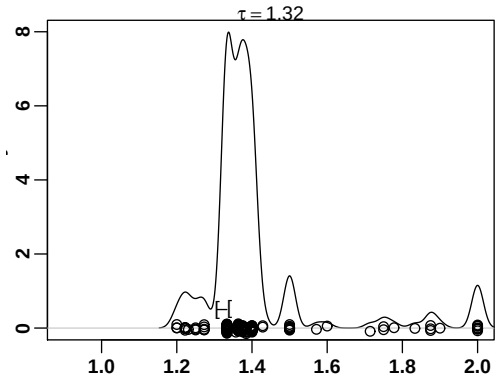

Figure 7.9: Kernel density estimate of the distribution of the Fourier estimator.

7.6.4 Linear Regression Estimator

As we noted in the introduction, it is natural to sort and index the observations in $\mathcal{S}$. An estimator for the overdispersion parameter τ can be deduced from the indexes by re-writing the observation model $X = \lfloor \tau N \rfloor$ into $X = \tau N + \varepsilon$ where ε is an error term. Even if ε is not Gaussian, linear regression can yield a reasonable estimate of the regression coefficient τ as Figure 7.10 shows.

We use the following algorithm:

- compute the empirical lattice $\{x_i\}$ by sorting and removing duplicates in the dataset
- compute the indexes $\{n_i\}$ according to the sorting index
- fit a regression line of the form $x_i = an_i + 0.5$
- return a

Figure 7.10 shows the linear regression estimator on the dataset $\mathcal{S} = \lfloor \tau [\![1, 10]\!] \rfloor$ (no missing values). For each element in the dataset, if the regression line intersects the length 1 interval then the estimate is compatible with the data point.

Note that the truncation error is not centered. Consequently, we compute the regression coefficient in the the model $X + 0.5 = \tau N + \varepsilon$. For the same reason, the regressors are below the regression line.

The main difficulty in the linear regression is that the values of the regressor variable N are unknown. In the distinguishable case, it is possible to reconstruct them when there are no missing values, i.e. $\mathcal{S}_\tau \cap [\![0, n]\!] = \mathcal{S}$ where $n = \max \mathcal{S}$. Otherwise, the regressors will be shifted and that affects strongly the estimate. Figure 7.11 shows such a case. The regressors inferred in the linear

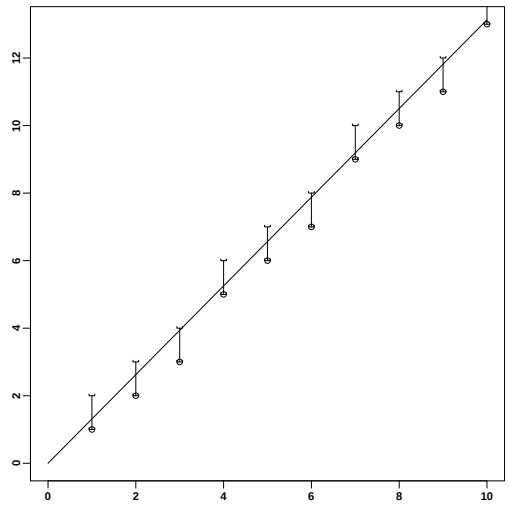

Figure 7.10: The linear regression estimator on a dataset without missing values.

regression estimator and the regression line are shown in solid line. For comparison, the true regressors are displayed in dotted line. The compatible values estimator finds the true regressors and its regression line is shown in dotted line.

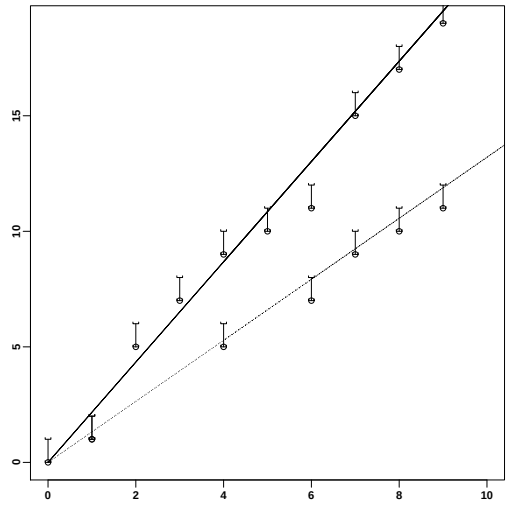

Figure 7.11: The linear regression estimator in the case of missing values. The compatible values estimator is shown in dotted line.

7.6.5 Discussion

In this section, we have compared the compatible values estimator with other estimators that correspond to different point of views on the problem. As we have shown in Proposition 7.5.4 that the compatible values estimator is optimal and that it extracts all the available information from the data set, we did not expect the alternative methods to outperform the CVE.

We checked that although the conditions in Proposition 7.5.4 are rather strict because they require a complete data set, in practice the estimator behaves well even in the presence of missing

values.

The alternative estimators are much worse and emphasize the difficulties related to not taking into account the discrete nature of the model. The worst option is to neglect missing values as is the case for the linear regression estimator. This estimator is very good on a complete data set, but is completely dumbfounded in the case of missing values.

The double Poisson family introduces additional knowledge in the form of a Poisson hypothesis on the Noise distribution. We showed that it is not suitable in the case of rare events, or when the quantization errors are significant. This is because the model is not adequate.

The Fourier estimator is very interesting because it shows that we can recover τ as a pseudo-period. A major difficulty is that 1 is an acceptable period, which is why 1.5 and 2 are observed in the estimator density in Figure 7.9. But the simulations show that τ can be recovered.

7.7 Extensions

In this section, we present extensions of the results that were not considered for publication in the manuscript reproduced in this chapter. Section 7.7.1 contains other mathematical results on the proposed model. Section 7.7.2 presents the application of the noise model with synthetic data.

7.7.1 Other mathematical results

7.7.1.1 Using holes in the data set

The results presented earlier only take into account positive information, i.e. the presence of integers in the data set. As a consequence, they only provide upper bounds on the gain parameter. The results in this section derive lower bounds on the gain parameter by considering negative information, i.e. the absence of some integers from the data set.

Recall Example 7.3.1 from Section 7:

```
> data = floor(1.32 * n)
% Distinguishible case
> n
 [1]  1  2  3  4  5  6  7  8  9 10
> data
 [1]  1  2  3  5  6  7  9 10 11 13
```

In this example, the integer 4 cannot be observed in a data set corresponding to $\tau = 1.32$. However, it is difficult in practice to distinguish integer values that do not belong to the true lattice (*non-observed values*) and integer values that are missing from the empirical lattice (*missing values*). This section provides some results based on non-observed values.

Definition 7.7.1. In the model $X = \lfloor \tau N \rfloor$, the *set of observed values* is the subset of the integers $D = \{x \in \mathbb{N} | \exists k \text{ and } x = \lfloor \tau k \rfloor\}$. The set of *non observed values* is the set $\mathbb{N} - D$. A *missing value* is an integer in D that is missing from the data set or empirical lattice.

Observed values provide upper bounds for τ. Conversely, lower bounds for τ can be deduced from non-observed values.

Proposition 7.7.2. *The set of observed values is $\mathbb{N}$ if and only if $\tau \leq 1$. Conversely, $\tau > 1$ if and only if there exists a non-observed value.*

Proof. • If $\tau \leq 1$. Let $k \in \mathbb{N}$ denote an integer, there exists $j \in \mathbb{N}$ such that $k \leq j\tau < k+1$. Consequently, $k = \lfloor j\tau \rfloor$ and k is in the set of observed values.
• If $\tau > 1$. Let $j \in \mathbb{N}$ such that $\tau - 1 > \frac{1}{j}$. Then $\tau j = j + (\tau - 1)j > j + 1$. In the interval $[\![0, j+1]\!]$, there are $j+2$ integers, but at most $j+1$ observed values.

□

More precisely, consecutive non-observed values can only appear when τ is large:

Proposition 7.7.3. *Let $i, j \in \mathbb{N}$. If all the integers in $[\![i, j]\!]$ are non-observed values, then $\tau > j - i + 1$.*

Proof. If the integers in $[\![i, j]\!]$ are non-observed values, then the lattice skips over the interval, i.e. there exists $j \in \mathbb{N}$ such that $\tau j < i < j + 1 \leq \tau(j+1)$. □

When τ is small, near 1, then we can combine observed values and non-observed values into the following proposition.

Proposition 7.7.1 (Maximal observed intervals)**.** *Suppose that $[\![x, y]\!]$ is a subset of $\mathcal{S}$, and additionally that $x - 1$ and $y + 1$ are non-observed values, that is to say $[\![x, y]\!]$ is maximal. Then we have the lower bound $\tau > 1 + \dfrac{1}{y - x + 2}$.*

Proof. • an interval of consecutive integers is observed, so according to 7.9.2, $\tau < 2$. Consequently, $x - 2$ and $y + 2$ must be observed values.
• let $x = \lfloor \tau n_x \rfloor$ and let $y = \lfloor \tau n_y \rfloor$
• $x - 2 = \lfloor \tau(n_x - 1) \rfloor$ and consequently $\tau(n_x - 1) < x - 1$
• $y + 2 = \lfloor \tau(n_y + 1) \rfloor$ and consequently $\tau(n_y + 1) \geq y + 2$
• by subtracting we obtain $y - x + 2 + 1 < \tau(n_y - n_x + 2)$
• this leads to $\tau > 1 + \frac{1}{y-x+2}$ because $y - x = n_y - n_x$ in the distinguishable case

□

7.7.1.2 Robustness

The compatible values estimator is very robust to missing values, and performs well on small data sets. With enough missing values, it may favor a value of τ that is too high. For example, when observing $\mathcal{S} = \{2, 4, 6\}$, the CVE returns 2 instead of $4/3$. The probability of such an event seems low, but is difficult to estimate in general.

The compatible values estimator is quite sensitive to outliers, i.e. to observations in the data set which belong to the set of non-observed values for a given τ. This can be alleviated by considering values that are compatible with the data set $\mathcal{S}$ with at most α errors.

The α-compatible values estimator is thus defined as the largest real number that is α-compatible with the data set $\mathcal{S}$. It can be found with the following algorithm, which is a slight modification of the algorithm for the compatible values estimator:
• compute the set of observed values by sorting the data set and removing multiple occurrences
• compute the upper bound $\tau < B = \dfrac{\hat{x} + 1}{\mathrm{card}\mathcal{S} - \alpha}$
• find an approximation of the largest compatible value t by testing the α-compatibility of the real numbers $t_k = B - \dfrac{k}{\hat{x}^2}$
• deduce the indexes from t, that is to say for all $x \in \mathcal{S}$, find i such that $x = \lfloor ti \rfloor$
• compute the interval $[a, b[$ as the intersection of the constraints $t \in \left[\dfrac{x}{i}, \dfrac{x+1}{i}\right[$, for all x in $\mathcal{S}$
• return $\dfrac{a+b}{2}$ as an estimator for τ

The algorithm is correct because of the following analogues to the results in Proposition 7.9.4 and Proposition 7.5.3:

Proposition 7.7.2 (Density Upper Bound). *Let τ denote an α-compatible value. Let $\hat{x} = \lfloor \tau \hat{n} \rfloor$ denote the largest integer in $\mathcal{S}$. Then $\tau < \dfrac{\hat{x}+1}{\hat{n}-\alpha}$. When $\hat{n}$ is unknown (because of potential missing values), let n denote the number of non zero observed integers. Then $\tau < \dfrac{\hat{x}+1}{\hat{n}-\alpha} \leq \dfrac{\hat{x}+1}{n-\alpha}$.*

Proposition 7.7.3. *The set of α-compatible values $\mathcal{C}(\mathcal{S})$ contains exactly $]0,1]$ and intervals of length at least $1/\hat{x}^2$ where $\hat{x} = \max \mathcal{S}$.*

Proof. Same proof as Lemma 7.5.1. □

7.7.2 Application to normalization

In this section we use the proposed noise model for spectrum normalization (see Section 6.8), and the estimation of the baseline. We use a synthetic LC/MS image in order to compare the estimated normalization factors (NFs) to the ground truth.

Synthetic LC/MS image The LC/MS image displayed in Figure 7.12 was generated based on a real mass spectrum in the following way:

- select one mass spectrum $\mathcal{I}_0(m)$ from a real LC/MS image (4-5 protein mix in [PMW+06]),
- create a LC/MS image by cloning the mass spectrum 500 times,
- add random Poisson noise $\mathcal{P}(1)$ with intensity 1 to every pixel,
- add a Gaussian signal PEP with retention time 320, m/z ratio 560 Da, standard deviation 7 and area under the curve 8000, as well as two isotopes,
- generate a vector NF_t of length 500 containing the true normalization factors (displayed in Figure 7.12),
- apply the normalization factors to each column of the LC/MS image.

We obtain the following LC/MS image:

$$\mathcal{I}(t,m) = \lfloor \mathrm{NF}_t \left(\mathcal{I}_0(m) + \mathcal{P}(1) + \mathrm{PEP} \right) \rfloor$$

Compatible Values Estimator (CVE) We use the Compatible Values Estimator defined in Section 7.5.3 to estimate the normalization factor in each mass spectrum $\hat{\mathrm{NF}}_t$. Figure 7.13 shows the results of normalization with the CVE, as well as a comparison with the true values. The CVE is very precise (optimal as demonstrated in 7.5.4) and the standard deviation of the relative difference $\hat{\mathrm{NF}}_t / \mathrm{NF}_t$ is 2.59×10^{-3}. The CVE is not influenced by the signal introduced in the image.

Total Ion Count estimator (TIC) We use the Total Ion Count estimator defined in Section 6.8.5. Figure 7.14 shows the results of normalization with the TIC. In particular, in the obtained LC/MS image, there is a vertical stripe at rt 320; this indicates that the TIC normalization factor was overestimated, as is shown in the lower panel. The TIC has a relative standard deviation of 13.6×10^{-3}, but is very sensitive to the peptide signal[5].

Quantile estimator We use the quantile estimator defined in Section 6.8.5. The quantile estimator is more robust than the TIC, but also less precise as shown in Figure 7.15; it has a relative standard deviation of 38.1×10^{-3}.

Figure 7.16 summarizes the performance of the three proposed normalization methods.

[5]This estimate was obtained after filtering the retention times corresponding to the peptide signal between 300 and 340. Without filtering, the relative standard deviation is 44.8×10^{-3}.

Remarks

1. The TIC and the quantile estimator cannot perform normalization without a reference. This reference is provided by applying each estimator to $\mathcal{I}_0(m) + 1$.
2. To remove bias from the quantization errors, we apply the TIC and the quantile estimator to $\mathcal{I}(t, m) + 0.5$.
3. The LC/MS image size was 500 scans times 400 pixels. The TIC estimator is nearly instantaneous. The quantile estimator takes on the order of several seconds. The CVE takes on the order of a minute, so it is not very computer intensive.

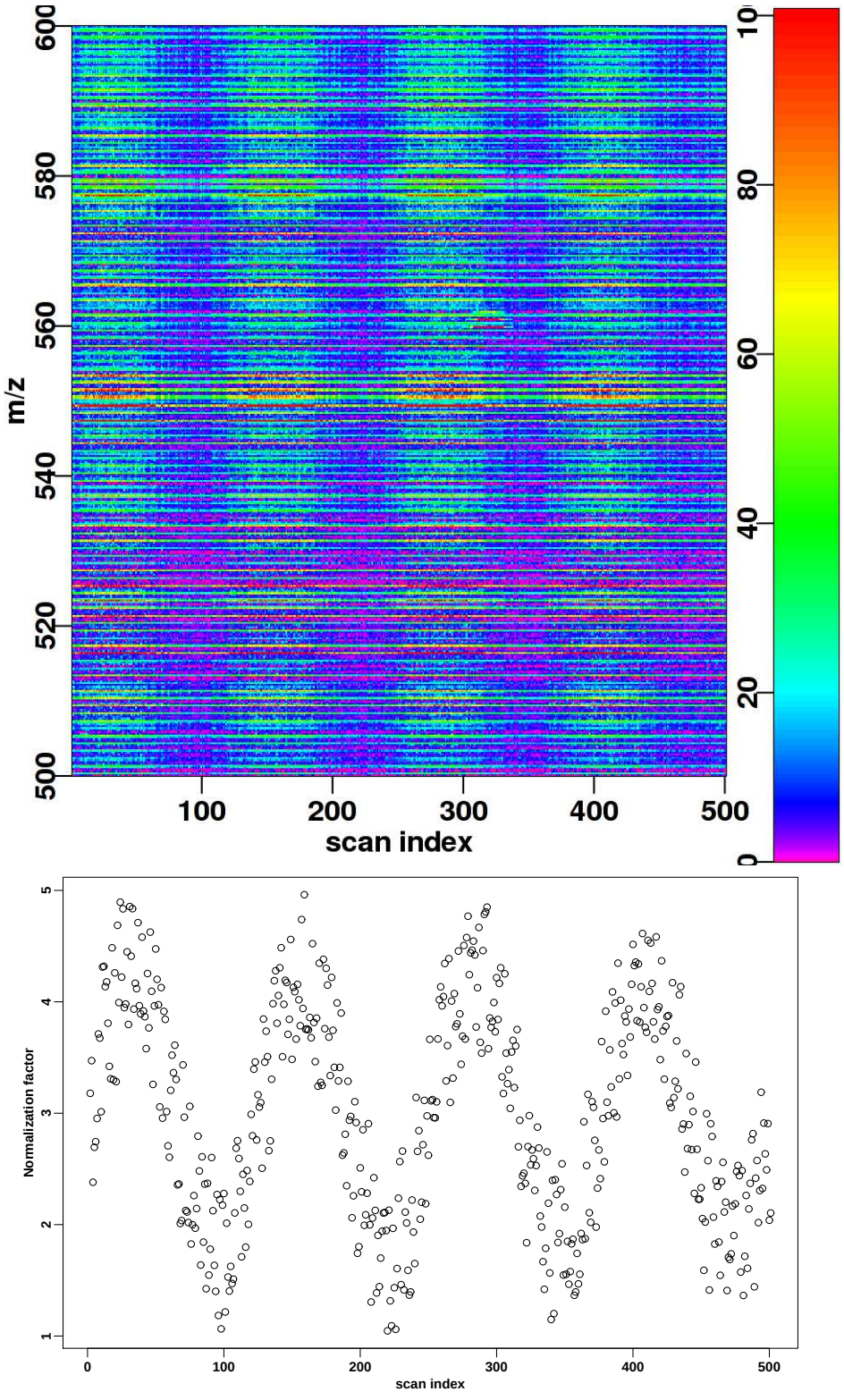

Figure 7.12: Synthetic data set. One mass spectrum was cloned to obtain the LC/MS image. We then added some noise and a Gaussian isotopic pattern at (rt=320,m/z=560). The lower panel shows the true normalization factors.

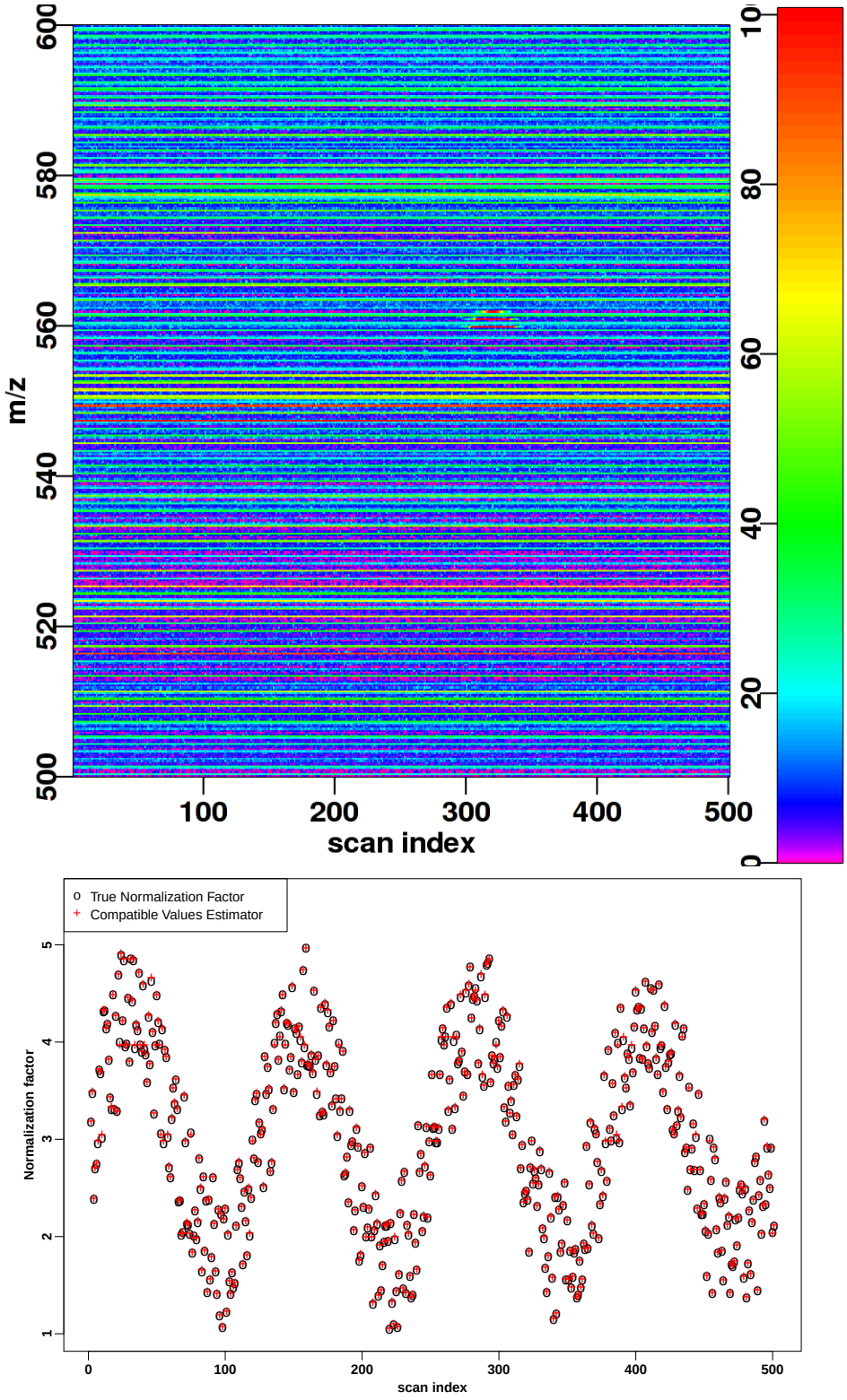

Figure 7.13: Normalization of the synthetic image in 7.12 with the Compatible Values Estimator. The lower panel shows the true normalization factors and the estimated normalization factors.

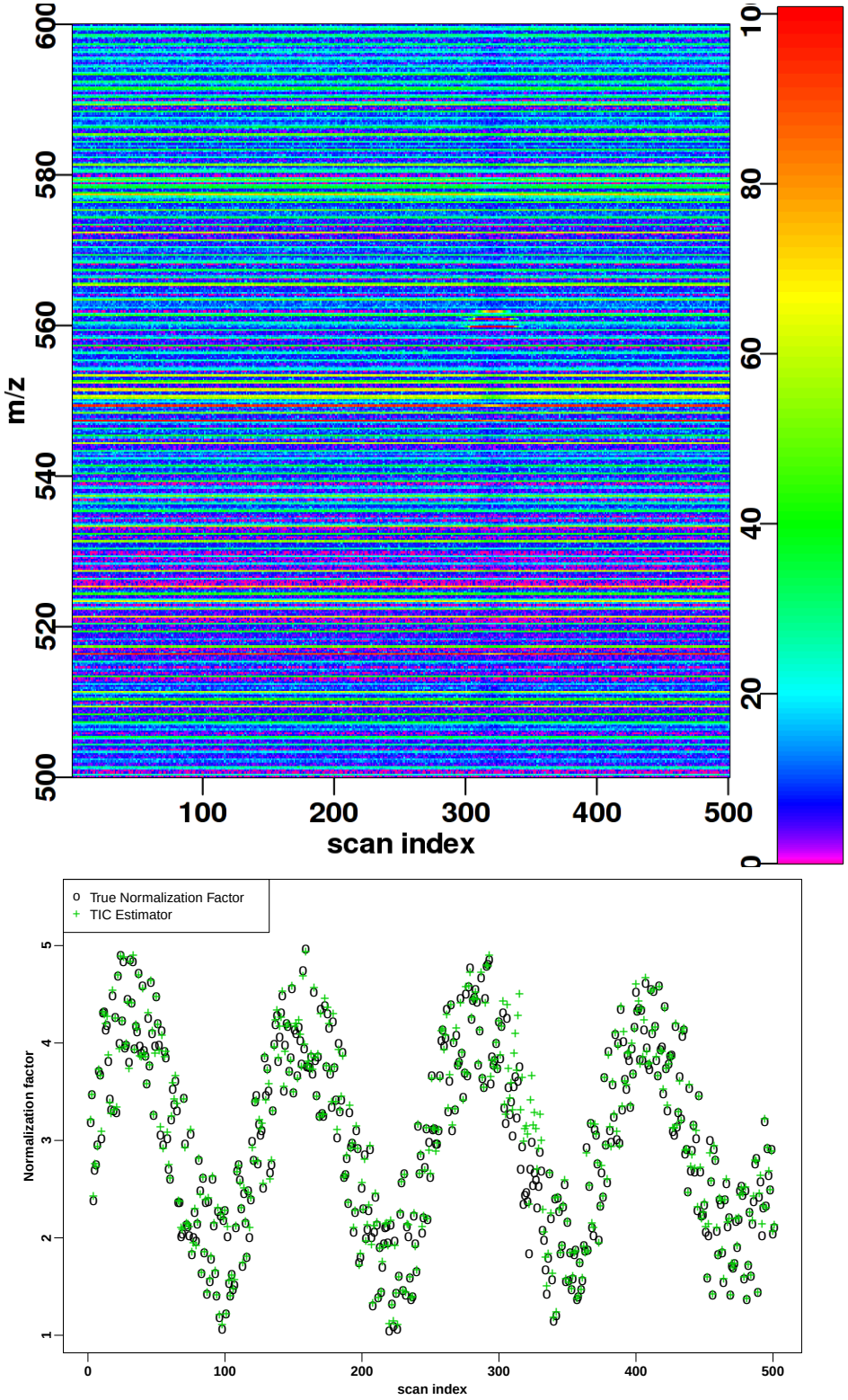

Figure 7.14: Normalization of the synthetic image in 7.12 with the Total Ion Count. The lower panel shows the true normalization factors and the estimated normalization factors. The TIC estimate is strongly influenced by the peptide signal at $(rt = 320)$, and creates a vertical stripe in the LC/MS image.

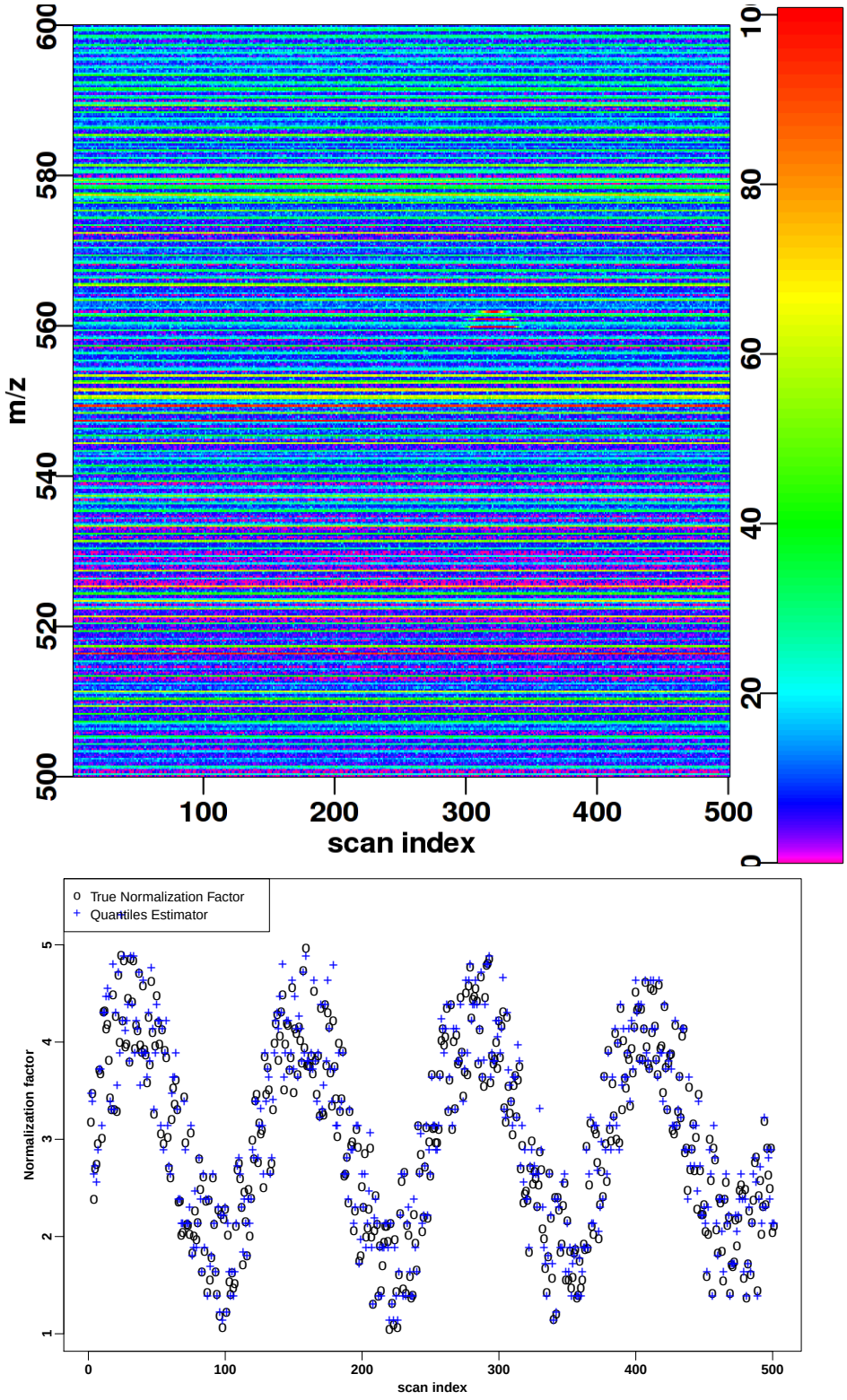

Figure 7.15: Normalization of the synthetic image in 7.12 with the quantiles estimator in 6.8.5. The lower panel shows the true normalization factors and the estimated normalization factors.

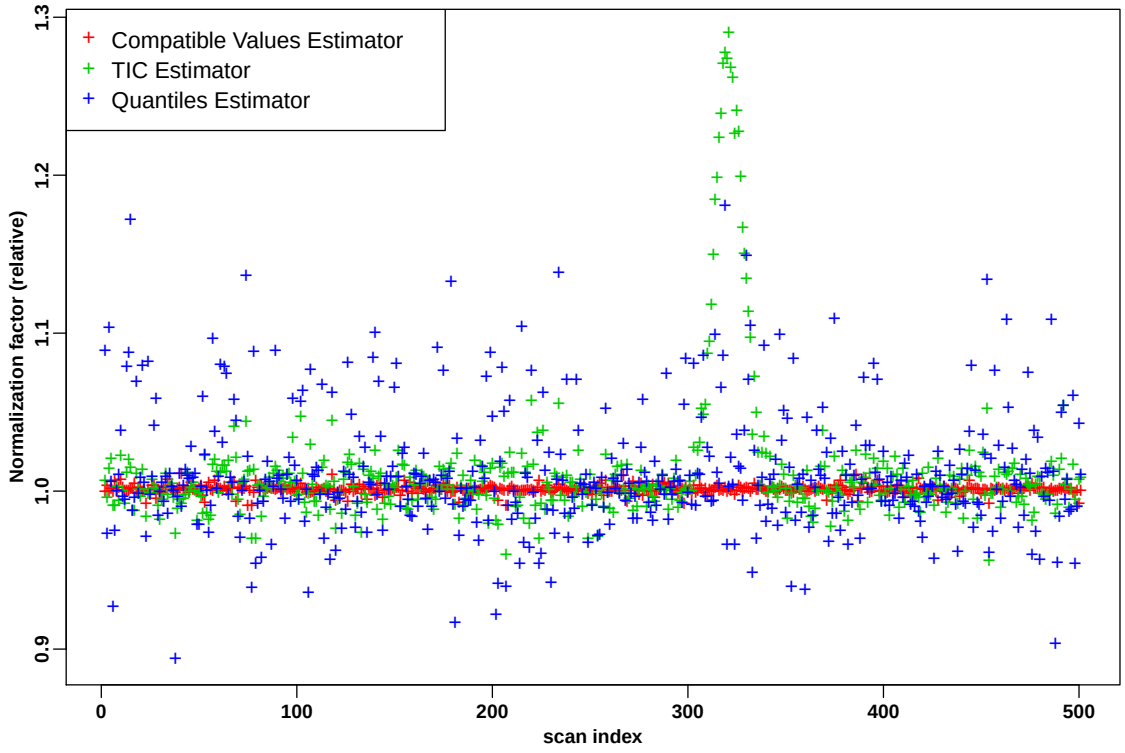

Figure 7.16: Comparison of the normalization factors obtained from the Compatible Values Estimator (CVE), the Total Ion Count (TIC), and the quantiles estimator. The CVE is an order of magnitude more precise than the others. The TIC is sensitive to the peptide signal at rt=320.

7.8 Related problems

Periodic signals As suggested by the Fourier estimator (Section 7.6.3), the problem of estimating τ is related to estimating the period of a signal ("harmonic retrieval problem"). The model presented in this chapter focuses on signals that are composed of periodic impulses, but with a high quantization error on the time of observation.

In this chapter, we suppose that the phase of the signal is equal to zero. When the origin of time is unknown, the model may be extended into the following:

$$X = \lfloor \tau N + \phi \rfloor$$

where ϕ is the phase of the signal.

We expect that bounds for τ in the extended model can be obtained with minor adjustments. However, the estimation procedure seems difficult to generalize.

Regular models From a statistical point of view, the proposed model is intriguing. First, it is not a regular model in the sense of [Mil01], but it is also not trivial due to the quantization error. It becomes a regular model — for a discrete variable — when considering an additional stochastic error term ε:

$$X = \lfloor \tau N + \varepsilon \rfloor$$

For the regular model, it would be interesting to use ε to model the dispersion of the kinetic energy of the ions in the mass spectrometer, and maximum likelihood estimation is a natural choice for regular models.

7.9 Conclusion

In the observation model $X = \lfloor \tau N \rfloor$, estimation of the overdispersion parameter τ and of the distribution of N can be decoupled. We have proposed an estimator for τ which allows the full recovery of the statistics of N prior to modeling. The structure of N may then be studied at length afterwards.

The estimator based on compatible values is optimal and quick. It is resistant to missing values in practice, and in the worst case returns an acceptable (parsimonious) answer without hypotheses on the law of N.

The compatible values estimator outperforms the other three estimators at the cost of a reasonable increase in computational load. For instance, the Double Poisson Family is not a suitable model in our range of parameters, but it may be preferable in large datasets and stronger Poisson-distributed Noise. Likewise, given a priori knowledge on the distribution of N, quantile regression could replace the linear regression estimator and improve the computation of the indexes.

The Fourier estimator suggests a strong relationship with the harmonic retrieval problem, despite the signal not being periodic. Although the truncation error considered in this paper is very different from Gaussian errors usually considered in harmonic retrieval, some algorithms from that field may make a better compromise between speed and precision for the current problem.

In Section 7.7, we propose extensions to the compatible values estimator that increase its robustness and apply the procedure to a synthetic data set to normalize the intensity. Unfortunately, the compatible values estimator only takes into account truncation noise, and yields poor results on real data. In Section 7.8 we propose to introduce an error term to better model real data and to introduce a phase parameter in the model to apply to a periodic signal context.

The noise distribution is not estimated in this chapter. Rather, we have proposed a normalization procedure that is independent of the noise distribution. In the next chapter, we continue on the same line and try to perform feature detection with as little hypotheses on the noise distribution as possible. We will only require that after normalization (by our method or another), the noise distribution is stationary in the retention time dimension, as observed in Chapter 6.

Appendix

Proposition 7.9.1. *The mapping $x \mapsto \lfloor \tau\, x \rfloor$ is injective if and only if $\tau \geq 1$.*

Proof. If $\tau = 1$, the mapping is the identity function. Suppose $\tau > 1$ and let n_1 and n_2 denote two (positive) integers such that $n_1 < n_2$. Then $\tau n_2 - \tau n_1 > \tau > 1$ and $\lfloor \tau n_2 \rfloor > \lfloor \tau n_1 \rfloor$. When $\tau < 1$, the mapping is not injective because $\lfloor \tau \times 1 \rfloor = \lfloor \tau \times 0 \rfloor = 0$. □

Upper Bounds on τ

Proposition 7.9.2 (Any two observations)**.** *Let x and y be two distinct elements of the set $\mathcal{S}$ of observed values. Then $\tau < 1 + |x - y|$.*

Proof. Let i and j be the values of N corresponding to x and y i.e. $x = \lfloor \tau i \rfloor$ and $y = \lfloor \tau j \rfloor$. Then we have the inequalities: $x \leq \tau i < x + 1$, $y \leq \tau j < y + 1$, and thus $\tau(j - i) < y - x + 1$. Assuming $x < y$, we obtain $\tau < \frac{y-x+1}{j-i} < y - x + 1$. □

Proposition 7.9.3 (Observed intervals)**.** *Let $[\![x, y]\!]$ denote the set of integers between x and y. If $[\![x, y]\!]$ is a subset of $\mathcal{S}$, then $\tau < 1 + \dfrac{1}{y - x}$.*

Proof. Using the same notations as in the proof of Proposition 7.9.2, $\tau < \frac{y-x+1}{j-i} < \frac{y-x}{j-i} + \frac{1}{j-i} < 1 + \frac{1}{j-i}$ because in the distinguishable case the number of elements in $[\![x, y]\!]$ is $y - x + 1 = j - i + 1$. □

Proposition 7.9.4 (Density Upper Bound)**.** *Let $\hat{x} = \lfloor \tau \hat{n} \rfloor$ denote the largest integer in $\mathcal{S}$. Then $\tau < \dfrac{\hat{x} + 1}{\hat{n}}$. When $\hat{n}$ is unknown (because of potential missing values), let n denote the number of non zero observed integers. Then $\tau < \dfrac{\hat{x} + 1}{\hat{n}} \leq \dfrac{\hat{x} + 1}{n}$.*

Chapter 8

Detection of protein signals in LC/MS images with the M-N rule

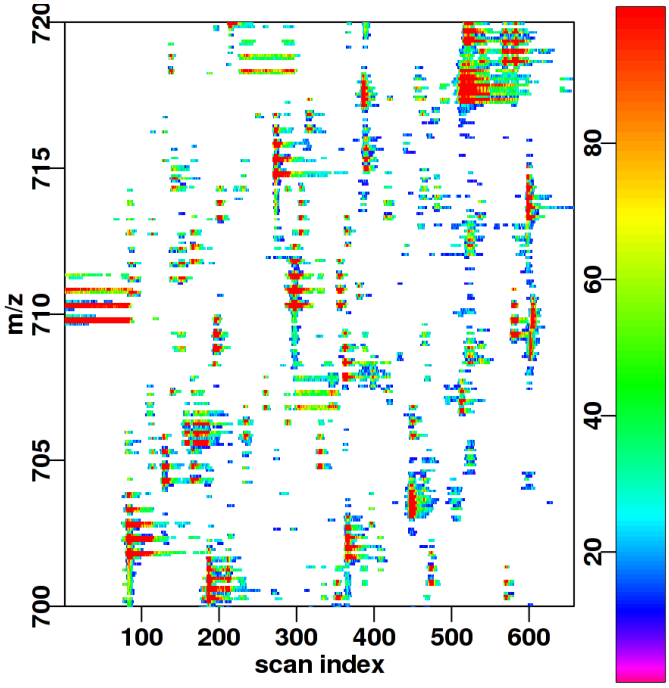

Figure 8.1: Feature detection in LC/MS images with the M-N rule.

8.1 Introduction

8.1.1 Context

As presented in Section 4.3, peptides in LC/MS images can be identified based on their m/z ratio and retention time. Before identification however, peptide signals must be detected in the image, and their m/z ratio and retention time must be computed. This chapter deals with the detection of peptides in a LC/MS image, and the content is extracted from [LTTS09].

Feature detection in LC/MS images corresponds to building a *list of locations* in the image where the measured signal intensity is attributed to the presence of a peptide.

We distinguish two separate steps. First, *candidate selection* determines regions of interest. Current approaches for candidate selection are mostly based on local maxima, either in the measured intensity [YPT+03, THN+04, NF07, KPRH05, MPP+07, WZP+06, YWL+06, KO05] or in the wavelet transform of the signal [RY06, LGR+06, BCF+06, TBN08, MCK+05, NF07]. In both cases, local maxima contain high numbers of false positives, and the list of candidates is filtered based on Signal-to-Noise ratio [MCK+05, NF07, MPP+07, WZP+06], peak width [YWL+06, KO05] or based on the reproducible presence of the peak in adjacent MS spectra [KPRH05, MPP+07, WZP+06]. The approach presented by [RJR+04], which we will here call the *Median M-N rule*, does not require the candidate to be a local maximum, but focuses on high intensity peaks that appear in several consecutive MS scans.

Candidate selection does not provide precise values for the signal characteristics. The second step, called *peak picking* is used for:

- determining the centroid of the signal, its retention time and m/z ratio
- computing the area under the curve, or other quantitative indicators
- the charge state of the ion
- the signal width in elution time and m/z
- the expected standard deviation of the centroid, etc.

The primary concern in peak picking is to accurately determine the m/z value of the peptide signal. To that end, most method match a template to the observed intensity values in the vicinity of the location obtained by candidate selection. This template matching usually requires more computationally-intensive procedures because of the optimization step (see [YHY09] for a survey of available methods and software).

There are several available models for individual peaks, including double Gaussian functions [KSBR04, SRS03, LSJ+06], asymmetric Lorentzian or sech functions [LGR+06] and exponentially-modified Gaussian functions [NH88, JXZ+08, Li02]. However, the template approach is best used to match the patterns of peaks created by isotopes [NF07, GMG+99, JPBF+04].

8.1.2 Why control the false positive rate ?

In the context of large-scale proteomics analyses, it is tempting to lower the requirements on the false positive rate of detection in order to increase the sensitivity [1]. As indicated in [DKL06], the performance of feature detection directly affects the subsequent processes, such as retention time alignment [Jef05], protein identification [RCA+04] and biomarker identification [LOW+05]. Only detected signals are identified and quantified.

A high false positive rate of detection is dangerous because the list of detected features is transmitted to the subsequent algorithms as ground truth and never again challenged. It leads to wrongly identified peptides, and in turn proteins. Due to the complexity of the signals and multiple sources of noise in MS spectra, high false positive peak identification rate is a major problem, especially in detecting peaks with low amplitudes [HKPM06].

[1] It statistics, this corresponds to increasing the test power at the expense of the false positive rate.

Feature detection algorithms are developed for LC/MS images rather than fragmentation spectra. In the latter, most of the chemical noise background is filtered by the selection of the parent ion m/z ratio. MS/MS spectra are much cleaner than their MS counterpart, and feature detection is considered an easy task. However, in LC/MS images, the challenge is on detecting low-intensity signals that correspond to low-abundance biomarkers, without too many false positive detections.

8.1.3 Summary

In this chapter, we study the candidate selection step because it drives the performance of feature detection in terms of sensitivity and selectivity. Our primary goal is to impose a statistical bound on the number of falsely detected signals in the LC/MS image. On the other hand, peak picking determines the precision of the later stages of the analysis (protein identification and quantification).

Among the various possibilities, we selected the Median M-N rule [RJR$^+$04]. This algorithm is computationally efficient and also amenable to statistical analysis. We show that the original formulation allows a limited level of control of the false positive rate. Therefore, we present an extended M-N rule in Section 8.2 and compute its statistical properties.

The extended M-N rule is placed in the framework of a contrario detection and statistical testing. The null hypothesis and the alternative are described based on models for the background noise and peptide signal in Section 8.3. The corresponding test statistic is presented in Section 8.4, and we compute the false positive rate in Section 8.4.1 and a lower bound for the false negative rate in Section 8.4.2. This bound is interpreted in terms of the limit of detection of the algorithm and in terms of a Signal-to-Noise ratio.

The theoretical results indicate how to set the detector parameters for optimal sensitivity, but the detector is adapted to only certain types of elution profiles. These claims are evaluated on a real data set in Section 8.5.

Section 8.6 discusses the validation of detection results. This is difficult in three aspects:

- real images contain low-intensity signals, which may be mistaken for false positives,
- standard evaluation methods do not take into account the variability of elution profiles,
- the notion of true and false positives is not clear because of the width of the elution profiles.

Two extensions to our work are presented in Section 8.7. We first study resampling of the LC/MS image, and show that an optimal pixel height can be chosen. Then we standardize elution profiles based on a relationship between their retention time and their width.

Finally, Section 8.8 puts the M-N rule in a wider context, and discusses related approaches in linear filtering, non-linear filtering and mathematical morphology.

The method presented in this chapter does not explicitly use the noise model presented in Chapter 7. Nevertheless, it is based on the hypothesis that the background noise distribution is independent of retention time after normalization as suggested in Section 6.8. Chapter 7 proposes such a normalization method, but other methods may be applied instead. On top of that, the noise model presented in Chapter 7 operates inside a mass spectrum (vertical line) whereas our feature detection method operates in the chromatography direction (horizontal line).

8.2 Feature detection

8.2.1 Feature detection with the M-N rule

In [RJR+04], the authors define the following feature detection algorithm. A peptide signal is detected by the Median M-N rule if the intensity in N consecutive MS spectra exceeds the threshold $M \times C$ where C is 30% of the trimmed mean or the median. The authors claim that the parameters $(M = 3, N = 3)$ can be used in many different contexts with low false positive rate.

When applying the Median M-N rule, we have observed that the false positive rate may be dependent on the m/z value, as is the case in Figure 8.2. This suggests that the false positive rate in the Median M-N rule is not well controlled. Consequently, the algorithm may provide an unspecified number of undesirable entries in the peak list, which may result in false identifications and wrong biomarkers. We defer the details of the estimation of the false positive rate to Section 8.5.4 because it builds on concepts and hypotheses provided in the rest of this section.

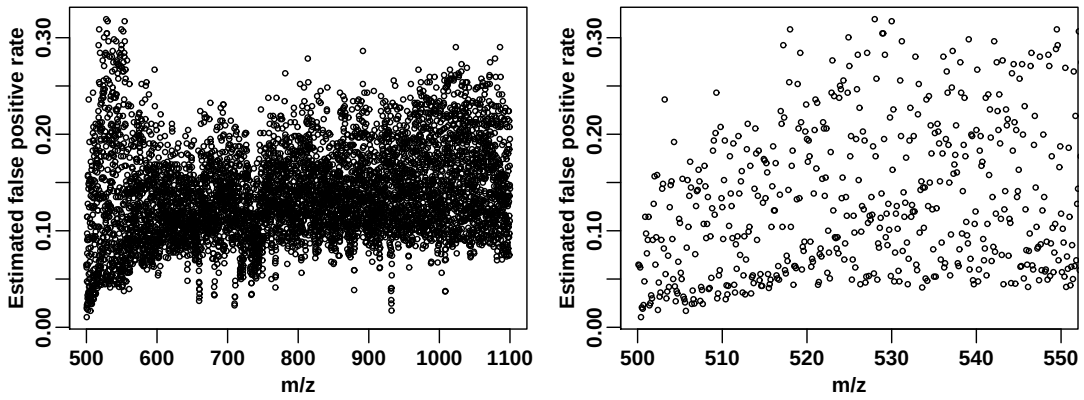

Figure 8.2: False positive rate in [RJR+04]. The data set and the estimation procedure are described in Section 8.5.1 and Section 8.5.4. The right panel is a zoomed portion of the m/z axis.

8.2.2 Extended M-N rule

We propose the following generalization of the M-N rule. A peptide is detected by the *extended M-N rule* if the intensity in N consecutive scans exceeds the threshold H. The *Median M-N rule* in [RJR+04] corresponds to using the threshold $H = M \times C$. In Section 8.4.1, we will present an alternative choice for H, which we call the *Quantile M-N rule*, and show how it improves the control of the false positive rate.

In [RJR+04], the Median M-N rule adapts to local noise characteristics although the parameters M and N are fixed for the entire data set. This is because the actual threshold $H = M \times C$ is a function of retention time and m/z through the median noise intensity $C(t, m)$. The extended formulation allows H to be an arbitrary function $H(t, m)$ of the position in the LC/MS image.

8.3 Generic Model of LC/MS Data

Computing the statistical properties of the detector requires a model of the signal generated on the LC/MS platform (described in this section) and procedures to estimate the model parameters (described in Section 8.5.2. We assume that the measured intensity $\mathcal{I}(t, m)$ is the sum of a random

noise component $\mathcal{N}(t, m)$ and of an independent and deterministic peptide signal component $\mathcal{S}(t, m)$.

Both the Median M-N rule and the extended M-N rule only take into account intensity in the same m/z bin, we will drop the variable m and write $\mathcal{I}(t) = \mathcal{N}(t) + \mathcal{S}(t)$. This is equivalent to analyzing the LC/MS data line by line.

As we analyze each m/z bin independently, a model for the m/z separation is not necessary. In the following, we describe the standard model for chromatography elution profiles presented in [SKG97], and the model for the background noise process $\mathcal{N}(t)$. To simplify the notations, we will hereafter write M-N rule instead of extended M-N rule, and study the generic properties of this feature detection algorithm.

8.3.1 Model for the Elution Profiles

In ideal chromatography experiments, each peptide produces a Gaussian-shaped elution profile ([SKG97]). This model has also been used to generate synthetic LC/MS images [STPG+08]. We assume that in each m/z bin, the signal $\mathcal{S}(t)$ is a superposition of Gaussian profiles, with different mean retention time and standard deviation. More explicitly:

$$\mathcal{I}(t) = \mathcal{N}(t) + \sum_k \Gamma_k(t)$$

where $\Gamma_k(t)$ is the Gaussian trace of peptide k in the current m/z bin, and the sum iterates over the peptides that have a trace in the bin.

Each Gaussian profile $\Gamma_k(t)$ is a positive real function of the chromatography time

$$\Gamma_k(t) = A_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(t-\mu_k)^2}{2\sigma_k^2}\right\}$$

where μ_k is the retention time of peptide k and the parameters σ_k and A_k represent respectively the standard deviation of the profile and its area under the curve. In particular, in LC/MS experiments, A_k is proportional to the concentration of peptide k in the sample.

In the standard model, the physical processes in distillation columns lead to the following relation between retention time and standard deviation:

$$\sigma_k = \mu_k / \sqrt{P} \tag{8.1}$$

where P is the plate number of the column. P measures the separation power of the chromatography; the elution profile is tighter (σ_k smaller) with increased values of P.

The plate number is commonly used to describe other types of columns, including those used in LC/MS experiments. However, it applies especially to isocratic mode. The use of solvent gradients may invalidate Equation (8.1) by modulating the retention of peptides on the LC column (μ_k) or the degree of separation (σ_k) during the course of the chromatography. Therefore, we do not use Equation (8.1) and we expect different possible values for σ_k.

8.3.2 Model for the Background Noise

Most methods dealing with background noise in LC/MS images attempt to remove it from the data using with various mathematical tools: wavelets [CTM+05, ZYZ03, QAT+03], Fourier transform [KGWR03], local noise statistics [SDM+04, WCD+05], time series analysis [ARC+03, LKP+03, HWC+03, ZWM+03, MCA+05]. However, removing background noise is difficult because chemical noise produces patterns that are similar to real signals [ARC+03]. For example, noise patterns have a 1 Da periodicity similar to isotope patterns.

To control the false positive rate in feature detection, we use the *a contrario* detection approach introduced in [DMM00, DMM01]. This approach is based on detecting image features as exceptional configurations in random images. As such, it requires that the noise characteristics be known a priori or estimated from the image. Several models for the background noise distribution in LC/MS images have previously been proposed [ARL+04, HNR02, WKG04, SMKM07] but are difficult to use in a contrario detection.

Instead of detailed modeling of the background noise, we use few hypotheses so that the approach remains generic. We simply assume that the random noise is an independent process and that all the pixels in the same horizontal line have the same distribution. Consequently we will write $\mathcal{N}$ instead of $\mathcal{N}(t)$. This requires that the LC/MS data is normalized beforehand, but greatly eases the practical estimation of local noise characteristics.

8.4 Statistical Testing Framework

Let us define a *pamphlet* of width N as a series of N intensity values $(\mathcal{I}(t_1, m), \ldots, \mathcal{I}(t_N, m))$ measured at the same m/z ratio m in consecutive MS spectra obtained at the retention times $(t_1, \ldots, t_N)$. In the following, we will use the notation $\mathcal{I}(t_i, m) = \mathcal{I}_i$ for all $i \in \{1, \ldots, N\}$. According to the hypotheses in the previous section, $(\mathcal{I}_1, \ldots, \mathcal{I}_N)$ are N independent random variables with distribution $\mathcal{N} + \mathcal{S}(t_i)$.

The models in Section 8.3 describe two different scenarios for the signal intensity in a given pamphlet. We say that the intensity follows the hypothesis H_0 when there is no peptide signal in the pamphlet. Conversely, the pamphlet follows the hypothesis H_1 when some peptide signals contribute to the intensity. This translates more precisely into:

H_0 : $\mathcal{I}_i = \mathcal{N}$ for all $i \in \{1, \ldots, N\}$, i.e. $\mathcal{S}(t) = 0$
H_1 : $\mathcal{I}_i = \mathcal{N} + \mathcal{S}(t_i)$ for all $i \in \{1, \ldots, N\}$ where $\mathcal{S}(t) > 0$

The M-N rule is a statistical test that decides whether a given pamphlet is under H_0 or H_1. It is iterated over all the possible pamphlets in the LC/MS image. In the following, we compute its false positive rate, i.e. the probability that the M-N rule detects a signal in a pamphlet under H_0, and then a lower bound on its power, i.e. the probability of not detecting a signal under H_1.

Distributions of the test statistic Given the parameter (H, N), the test statistic T is the sum of N Bernoulli random variables :

$$T = \sum_{i=1}^{N} \mathbb{1}_{I_i > H}$$

The M-N decision rule corresponds to $T \geq N$. Under H_0, T is a binomial random variable with parameters $(N, \mathbb{P}[\mathcal{N} > H])$. Under H_1, T is the sum of N Bernoulli random variables with parameters $\mathbb{P}[\mathcal{N} + \mathcal{S}(t_i) > H]$. In that case, the Bernoulli parameters of the N random variables are not the same.

8.4.1 Selectivity of the M-N rule

Given a pamphlet, the false positive rate α is the probability that the noise exceeds the threshold H in all N consecutive scans. Following the hypothesis that the noise is independent identically distributed we obtain :

$$\alpha = \mathbb{P}\left[\mathcal{N}(t_i) > H \text{ for all } i \in \{1, \ldots, N\}\right] = \mathbb{P}[\mathcal{N} > H]^N. \tag{8.2}$$

or equivalently

$$H = q_{\mathcal{N}, 1-\alpha^{1/N}} \tag{8.3}$$

where $q_{\mathcal{N}, 1-\alpha^{1/N}}$ is the quantile in the noise distribution of level $1 - \alpha^{1/N}$.

Quantile M-N rule In order to obtain tight control of the false positive rate α, we suggest that α should be chosen by the user and the local threshold $H(t, m)$ be derived from Equation (8.3). This choice of H is optimal because higher values are unnecessarily conservative while lower values increase the false positive rate beyond the user choice α. We call *Quantile M-N rule* the instantiation of the extended M-N rule where $H = q_{\mathcal{N}, 1-\alpha^{1/N}}$.

By Equation (8.3), the value for the threshold H in the Quantile M-N rule depends on the number of consecutive scans N. Increasing the value of N yields a lower threshold H, i.e. we can relax the threshold condition while maintaining the same false positive rate. This improves the detection of low abundance signals, but only under certain conditions, as discussed in the next section.

Multiple tests To detect all the peptide signals in an LC/MS image, the M-N rule is iterated over all possible pamphlets. This is a multiple testing scenario and we expect a relatively large number of false detections overall depending on the choice of α. To control the overall false positive rate, Bonferroni correction selects a very strict value for α such that α is the probability of obtaining one false positive in the whole image, leading to a very small number of detected peptides. Instead, we bound the number of false alarms NFA (cf [DMM00]), i.e. the expected number of false positives in the whole image.

$$NFA = \alpha \times width \times height$$

As the NFA is proportional to α, the user can set either.

8.4.2 Sensitivity of the M-N Rule

The sensitivity (test power) of the M-N rule is the ability to detect a peptide in a given pamphlet, i.e. it is the probability that $T = N$ under H_1. Contrary to the false positive rate which only depends on the noise characteristics, the sensitivity also depends on the actual shape of the signals in the pamphlet. A high intensity signal is easier to detect. In the following, we study the limit of detection as a function of signal shape, i.e. we try to determine the shapes with lowest area that can be detected reliably.

With most noise processes, any shape is detectable with a non-zero probability if the false-negative rate of detection is unrestricted (potentially high). Consequently, we focus on shapes that are detected with probability at least $1 - \beta$, with $\beta > 0$ a user set parameter. In the statistical testing framework, this corresponds to shapes for which the test has power or sensitivity above $1 - \beta$. This corresponds to the IUPAC recommendations, as stated in [Cur99].

As described in Section 8.3, we model the elution profile of a signal with a Gaussian function that is characterized by three numbers: the retention time μ, the standard deviation σ and the area under the curve A. Detectability does not depend on μ because we obtain an equivalent situation by translation, so we suppose from now on that $\mu = 0$. We suppose that there is only a single peptide signal in the sliding window. This corresponds to the following intensity :

$$\mathcal{I}(t) = \mathcal{N}(t) + A\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{t^2}{2\sigma^2}\right\}$$

With the same notations as in the previous section, we solve

$$\mathbb{P}[\mathcal{I}_i > H, \forall i \in \{1, \dots, N\}] > 1 - \beta$$

which leads to (details in the appendix)

$$A\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{N^2/4}{2\sigma^2}\right\} > H - q_{\mathcal{N},1-(1-\beta)^{1/N}} \tag{8.4}$$

or equivalently

$$\frac{A}{H - q_{\mathcal{N},1-(1-\beta)^{1/N}}} > \sqrt{2\pi\sigma^2}\exp\left\{\frac{N^2/4}{2\sigma^2}\right\} \tag{8.5}$$

This is a conservative approximation. Peptide signals with area A and standard deviation σ are guaranteed to be detected with a probability greater than $1 - \beta$ if Equation (8.4) is verified. Peptides with lower area can still be detected, albeit with lower probability.

The right-hand side of Equation (8.5), $F = \sqrt{2\pi\sigma^2}\exp\left\{\frac{N^2/4}{2\sigma^2}\right\}$ is independent of the threshold H. Consequently, we can optimize the choice of the parameters H and N independently. In practice, we suggest to choose N according to the typical extent of elution profiles in the LC/MS image, then compute the optimal choice $H = q_{\mathcal{N},1-\alpha^{1/N}}$.

Equation (8.5) leads to the definition of the Signal-to-Noise ratio as $\frac{A}{q_{\mathcal{N},1-\alpha^{1/N}} - q_{\mathcal{N},1-(1-\beta)^{1/N}}}$. In comparison with the classical ratio (maximum / standard deviation), this ratio suits better our intuition. First, the intensity of the signal is the area under the curve which relates better to the notion of low-concentration peptides than the maximum of the peak. Second, one quantile is controlled by the sensitivity parameter α and the other is independently controlled by the test power $(1 - \beta)$. It is easy to understand how the parameters affect the efficiency of the feature detection scheme.

F expresses how well the detector is suited to the shape. Detection is easiest when F is minimal, i.e. when $\sigma = N/2$. However, the converse is not true, and given σ, the detector with parameter $N/2$ may not be optimal as shown on Figure 8.3.

F does not depend on the noise distribution, but only on the Gaussian shape of the signals. As a consequence, a given detector is suited to a range of shapes that is roughly independent of noise distribution. More precisely, the two quantiles of the noise distribution $q_{\mathcal{N},1-\alpha^{1/N}}$ and $q_{\mathcal{N},1-(1-\beta)^{1/N}}$ affect the lower limit of detection, but not the adequacy between σ and N. In particular, the simulations presented on Figure 8.3 are expected to accurately reflect the real detection range in LC/MS images.

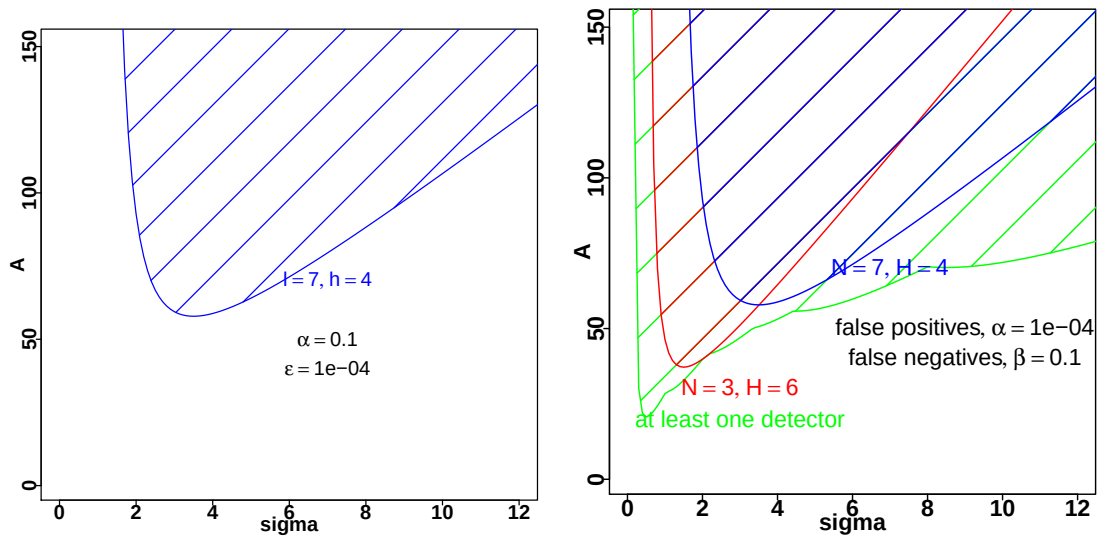

Figure 8.3: Limit of detection of the M-N rule. The shading lines indicate the set of shapes (A, σ) that a given detector can recall with probability greater than 90%. We show in green the fundamental limits of the M-N rule, i.e. the union of all sets for all possible values of the parameters.

In the following simulations, we assume that $\mathcal{N}$ is a Poisson noise process with mean 3, which is coherent with our experimental data set. In Figure 8.3, the left panel shows the set of shape parameters (A, σ) that the detector $(H = 4, N = 7)$ can recall with probability over 90%. The right panel, allows the comparison of the detectors $(H = 6, N = 3)$ in red and $(H = 4, N = 7)$ in blue, and also the union of the sets for all choices of H and N. The green set thus delineates the fundamental limits of the M-N rule with regard to Gaussian signals.

We observe that no single detector is universal because it is not suited to arbitrarily small values of σ or large values of σ. This motivates the use of several detectors with varying values of N to increase coverage. However, combining the sets of detections obtained by different algorithms is again a multiple testing problem and is outside the scope of the present paper.

Instead of using simulations, it is possible to estimate the background noise distribution in the LC/MS images and redraw the above diagram. This would be desirable to better adjust the parameters N and H. However, this is of no practical use because there are several background noise distributions in the LC/MS image.
Background noise is usually more intense in the middle of the mass range and there is a periodic component that corresponds to background chemical impurities (proteolytic background, contaminants, see [KSYW08] for a review). In particular, each horizontal line has a different noise distribution, if only because of the variations in the baseline component.

Conclusion The range of detected low-intensity peptide signals varies significantly with the parameters of the detector. However, there remain shapes that cannot be detected by the M-N rule regardless of the parameters because these are below noise level. Even when the noise distribution is not known, Figure 8.3 can provide generic guidelines to select adequate values for H and N.

8.4.3 Properties of the M-N rule

The feature detection algorithm presented in this chapter is generic and can be applied in many contexts. While we use normalized data to facilitate estimation of the noise level and its quantiles, only the hypothesis that noise intensities are independent is compulsory for the bound on the false positive rate in Equation (8.4). Likewise, it can be applied to centroided data, although we expect the centroiding to affect the estimation of the noise quantiles. It guarantees a uniform false positive rate in the LC/MS image, but also between images.

Independence is a critical assumption, therefore the extended M-N rule is best applied to raw data, without prior smoothing or baseline removal. The baseline is mostly harmless and it is taken care of in the estimation of the local noise quantiles. However, smoothing introduces correlation between adjacent pixel values and may lead to a higher false positive rate than anticipated.

We analyzed the performance of the detector is presented with a Gaussian model for the peak elution profile. Although this seems restrictive, it is useful and easy to interpret in terms of a Signal-to-Noise ratio. Moreover, the detected features are not restricted to Gaussian signals and can address other (e.g. asymmetric) types of signal. The Signal-to-Noise ratio defined in Equation (8.5) remains valid on condition that the elution profile width at height H matches that of a Gaussian function.

8.5 Results

8.5.1 Description of the Data Set

We use the data set published in [KEH+08]. A mix of 18 proteins was prepared and run on several mass spectrometry platforms including different types of instruments and replicates. The complete list of proteins in the standard sample is available, along with the experimental procedures used on each platform.

To evaluate the M-N rule feature detection algorithm, we focus on data acquired on a Q-TOF instrument. The illustrations were generated using the file `QT20060926_mix4_19.mzXML` from `mix4`. This instrument has sufficient resolution in order to distinguish isotope patterns, and the data is acquired in profile mode rather than centroid mode.

Centroiding is a signal processing algorithm that reduces the complexity and size of the data set. In doing so, it strongly affects MS signals and our binning procedure by shifting the position of MS data points. Moreover it affects the background noise distribution by aggregating peaks; in particular, we can observe empty noise regions alongside high intensity peaks in centroided data. Centroiding is usually selected as a default option in manufacturer software and cannot be undone.

We expect our algorithm to detect lower intensity peptides on a non-centroided data set because the noise distribution will be more faithfully estimated. For example, in regions of low noise, low intensity peptides can be more easily detected. In regions of higher noise, the false positive rate is better controlled. Because of centroiding, the estimate of the noise distribution may be incorrect. Centroiding can also lead to unnecessarily conservative choice for the parameter H and thus less detections than without centroiding.

8.5.2 Feature detection

Preprocessing The experimental data was loaded from the mzXML data file and subsampled into pixels of width one scan and height 0.1Da. This is similar to binning each mass spectrum with bin width 0.1Da. We chose to set the intensity of each pixel to the sum or integral of the intensities of the peak that belong inside the pixel. The background noise distribution appears uniform only in subregions of the data set. We chose to crop the data set to a retention time range rather than apply a normalization tool to avoid potential biases, and consider the cropped LC/MS image as normalized.

We chose a narrow mass bin width of 0.1Da because at that resolution we can observe a 1Da periodicity in the noise distribution and take it into account in the estimate of the parameter H. The periodicity is related to the background noise being chemical noise, i.e. random fragments of molecules including peptides. At the same time, the QTOF is an instrument with mass precision on the order of 50 ppm and provides enough measurements in each pixel for accurate estimation of the noise quantiles.

Noise quantiles For each m/z bin, we compute the threshold $H = q_{\mathcal{N},1-\alpha^{1/N}}$ from a local estimate of the quantile of the noise distribution. As the data is normalized, we compute the quantile using the straightforward empirical quantile from all the pixels in the mass bin. This procedure is quick and provides one threshold per line. For unnormalized data, we advise to use neighboring pixels in the mass bin to estimate the noise quantile.

The empirical quantile provides an estimate of H that is unbiased when there are no peptide signals in the m/z bin, and conservative in the presence of peptides. This is because peptide signals can only increase the quantile $q_{\mathcal{N},1-\alpha^{1/N}}$. The false positive rate always complies with the user set parameter α, regardless of the presence of peptides, and uniformly in the LC/MS image.

Feature detection In this context, a pamphlet is a horizontal series of N pixels that corresponds to the binned intensity values in consecutive scans. According the results in Section 8.4.1, a pamphlet contains a peptide signal if all the pixel intensities are above the threshold H. In the figures, when a pamphlet is detected, all the pixels belonging to the pamphlet are displayed.

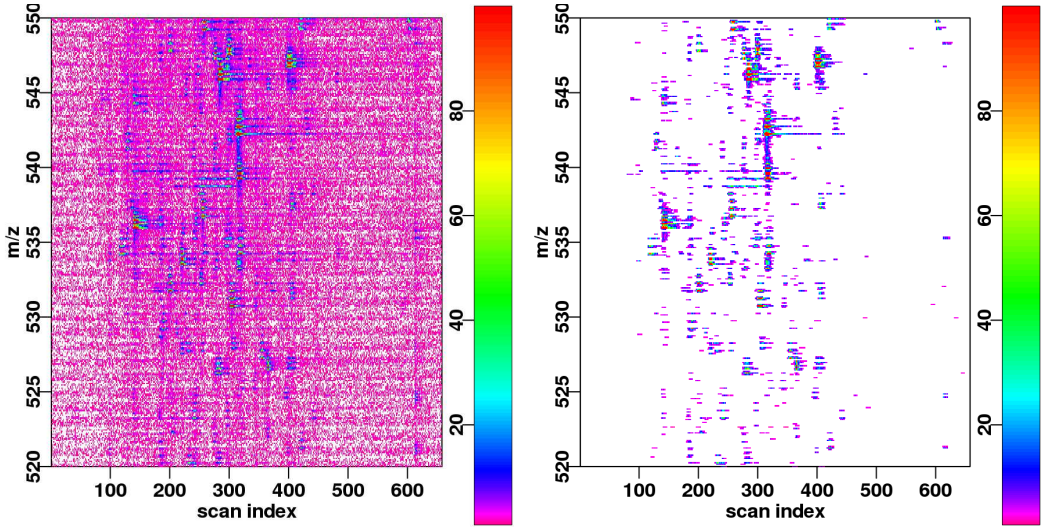

Figure 8.4: Detection results with $N = 7$ and $H = q_{\mathcal{N},1-\alpha^{1/N}}$. The left panel shows the raw image and the right panel shows only the areas where a peptide is detected.

Image results In Figure 8.4, we show an example of detection with $\epsilon = 10^{-3}, N = 7$. We observe that the detector can recall low-abundance signals with low error-rate. As a sanity check, it recalls isotope patterns without a priori knowledge.
We also verify that the detected signals do not display the 1Da period behavior of the background noise. This means that the false positive rate is independent of the changes in noise distribution at different m/z values.

In Figure 8.5, we show that α effectively controls the false positive rate by displaying the detection results in a relatively empty region of the image for several values for α. The user can set the false positive rate regardless of the noise level and the detector adapts automatically by adjusting the threshold H. With higher false positive rate, the detector is able to recall lower intensity signals, which are not significant under stricter constraints.

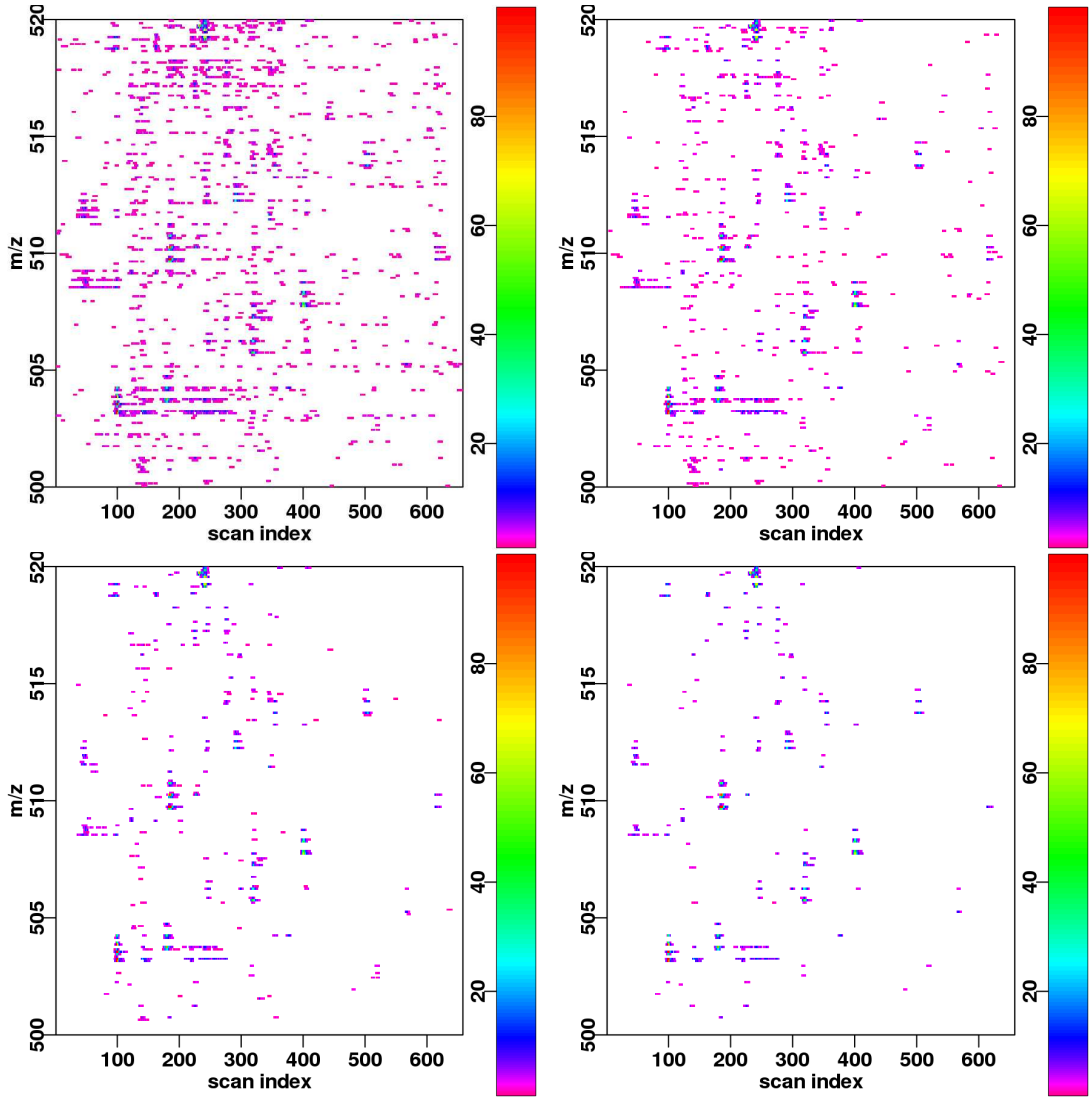

Figure 8.5: Detection results with varying false positive rate $\alpha \in \{0.1, 0.01, 0.001, 0.0001\}$. The number of false alarms per line in the images is respectively 65.7, 6.57, 0.657, and 0.0657. Bonferroni correction at level 5% would require $\alpha = 2.54 \times 10^{-7}$.

8.5.3 Adjusting Performance to Signal Shape

As emphasized in Section 8.4.2, the detector's performance is dependent on the signal shape. In this section, we compare three choices of the detector parameters $l \in 3, 7, 17$. We first display the theoretical range of the detectors in Figure 8.6.

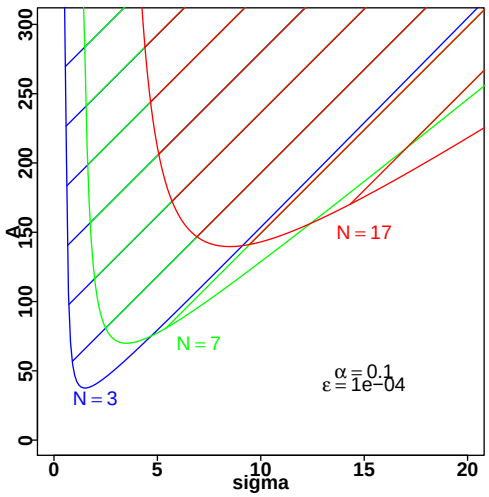

Figure 8.6: Theoretical detection range

The selected detectors are suited to different ranges of σ. As a consequence, they do not detect the same peptide signals in the LC/MS image as shown in Figure 8.7. The $(N = 3)$ detector is able to recall an isotopic pattern at $(rt = 110, m/z = 707Da)$ which is missed by the $(N = 17)$ detector. Conversely, the $(N = 17)$ detector is able to recall signals of lower intensity at $(rt = 50, m/z = 710Da)$ and especially the tails of the elution profiles. In this LC/MS image, the $(N = 7)$ detector seems to be a good compromise.

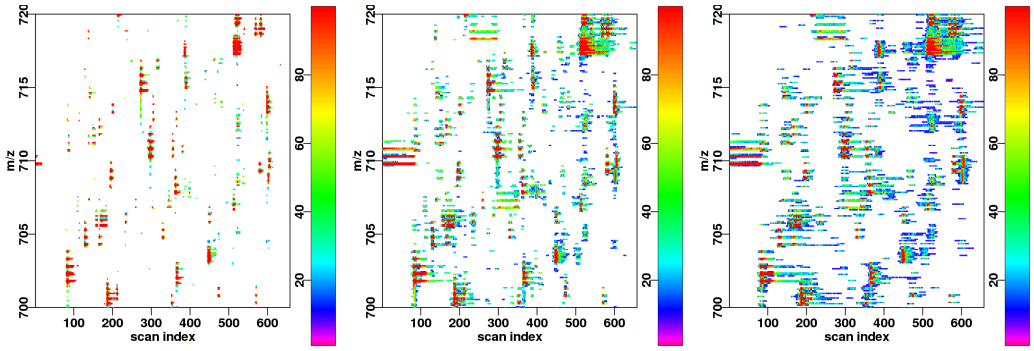

Figure 8.7: Detection results with length parameter $N \in \{3, 7, 17\}$

One drawback of the M-N rule is that some peptide may be randomly split. This occurs when a pixel intensity belonging to a faint peptide signal is lower than the threshold. The peptide can still be detected, identified and quantified thanks to the adjacent pixels, as can be seen in the close-up in Figure 8.8.

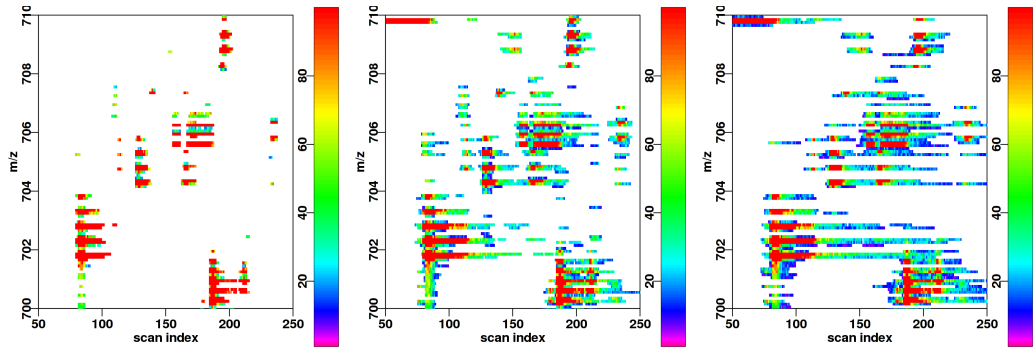

Figure 8.8: Detection results with length parameter $N \in \{3, 7, 17\}$

As indicated in Section 8.4.2, the M-N rule detector with parameters (N, H) is best suited for detecting Gaussian shapes of standard deviation $\sigma = N/2$. For other values of σ, the limit of detection is higher (increased area under the curve A). A reasonable heuristic for choosing N in practice thus consists in setting $N = 2\hat{\sigma}$ where $\hat{\sigma}$ is the mean width of the observed elution profiles in the LC/MS image. Note that in gradient elution, the liquid chromatography protocol can be adjusted so that the standard deviations of the elution profiles are roughly the same.

8.5.4 Comparison with the original M-N rule

In [RJR+04], the authors propose to use the Median M-N rule with the parameters $(N = 3, H = 3 \times C$ where C is 30% of the trimmed mean or the median. In this section, we discuss the advantages and the drawbacks of our proposed extension.

We compare the following two detection algorithms:

- Quantile M-N rule with $N = 3$, $H = q_{\mathcal{N},1-\alpha^{1/N}}$ and $\alpha = 10^{-1}$, and
- Median M-N rule with $N = 3$, $H = 3 \times C$ where C is the trimmed mean

Both C and $q_{\mathcal{N},1-\alpha^{1/N}}$ are computed using the pixel intensities in the current m/z bin. We choose trimming rather than the median because the data file contains integer-valued intensities instead of floating-point intensities. As the level of trimming is unspecified in [RJR+04], we chose to use symmetric 10% trimming.

In the original publication, the Median M-N rule is used after smoothing the intensities. This preprocessing step is not necessary for studying the properties of the M-N rule and is thus left out in the present paper. Smoothing can be applied before both algorithms with the following caveat. Smoothing reduces the noise variance, but introduces correlation between adjacent pixel intensities. As the pixel intensities are no longer independent the false positive rate computation in Section 8.4.1 is not valid. Consequently, smoothing may introduce artifacts in the detections.

As seen in Section 8.4, the choice $H = M \times C$ in the Median M-N rule is not generic. It provides a uniform false positive rate when the background noise distribution is Gaussian with mean C and standard deviation C, in which case $3 \times C = q_{\mathcal{N},0.977}$. However, the false positive rate is not controlled in the presence of baseline variations, which affect the mean background noise intensity but not its standard deviation. Alternatively, if the background noise is modeled as Poisson shot noise with mean λ, then its standard deviation is $\sqrt{\lambda}$, so $H = M \times C$ is too conservative with high noise intensity. In contrast, setting $H = q_{\mathcal{N},1-\alpha^{1/N}}$ ensures the same false positive rate in all situations.

On a real data set, we can illustrate the situation by deriving a theoretical value for the false positive rate based on the local threshold used by each algorithm. For each algorithm, we compute the threshold $H(m)$ in each mass bin, and use Equation (8.2) to obtain the predicted false positive rate. In Figure 8.9, we observe that the computed false positive rate for the Median M-N rule is very high and very variable as a function of m/z. On the other hand, the false positive rate of the extended M-N rule obeys the selected bound α. The variations of the false positive rate are due to quantization in TOF data.

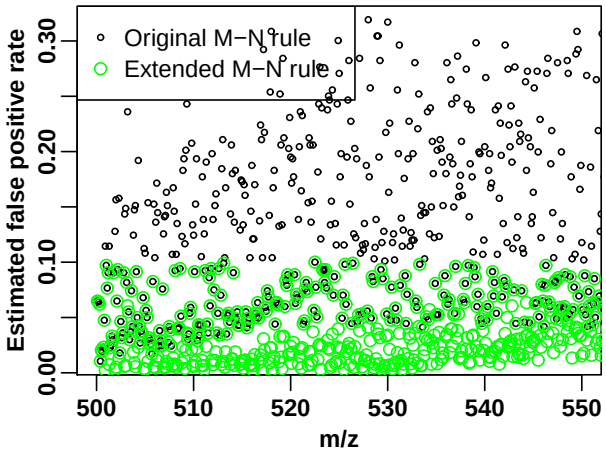

Figure 8.9: False positive rate as a function of m/z for the Median M-N rule and the Quantile M-N rule with $\alpha = 0.1$. In the extension, the false positive rate is variable but obeys the upper bound.

In Figure 8.10, we show the detection results obtained by the two algorithms on the same LC/MS image. In particular, we observe a periodic pattern with the Median M-N rule that corresponds to the periodic behavior of the noise. In both cases, the true signals are adequately detected.

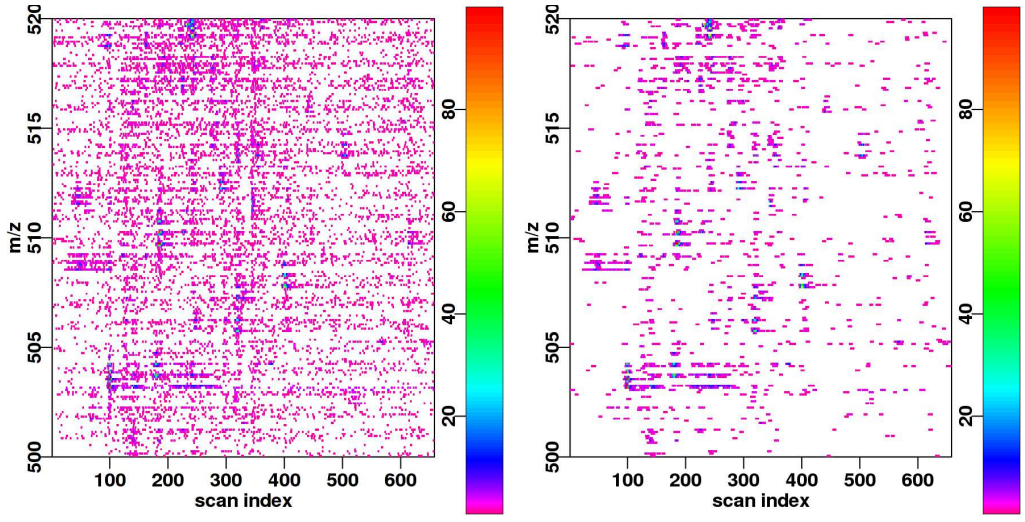

Figure 8.10: Detection with the Median M-N rule with parameters $(N = 3, M = 3)$ and the Quantile M-N rule with parameters $(N = 3, H = q_{\mathcal{N},1-\alpha^{1/N}})$ where $\alpha = 0.1$.

8.6 Validation of detection results

8.6.1 Visual evaluation on real images

Feature detection algorithms are evaluated based on the following criteria:

- false positives,
- false negatives,
- localization, i.e. the precision of the position in the LC/MS image,
- feature characteristics such as area under the curve, profile width,
- tolerance to noise,
- tolerance to departure from the feature model.

In this chapter, we have focused on the candidate selection step and thus on false positives and false negatives. Localization and other feature characteristics are left for the pattern matching step to determine.

The proposed method is tolerant to noise, but requires independence and stationarity, which are two properties that we expect in LC/MS images. The method is very tolerant to departure from the elution profile model; it can tackle non-Gaussian signals, with very different shapes.

Evaluation of feature detection based on visual assessment is possible, but is tricky in our context. If we use LC/MS images from large-scale proteomics experiments (which is the target application), we expect large numbers of low-intensity signals. Given a detection in this complex image, it is not possible to distinguish between a false positive and a faint signal. Isotope patterns only help partially in this respect. They can be used to validate a detection when a full pattern is observed, but their absence cannot invalidate a detection because the isotopes of light molecules may be undetectable (see Figure A.2 in the Appendix). Even low-complexity mixtures of proteins like the 4-5 protein mix in [PPW+07] used to generate Figure 6.16 on page 117 contains large numbers of low-intensity signals.

8.6.2 Visual evaluation on synthetic images

As the ground truth is not available in proteomics data sets, we have generated synthetic LC/MS images. We generate an independent identically distributed background noise $\mathcal{N}(t, m)$ and Gaussian signals as shown in Figure 8.11. These two components are added to form the image in Figure 8.12.

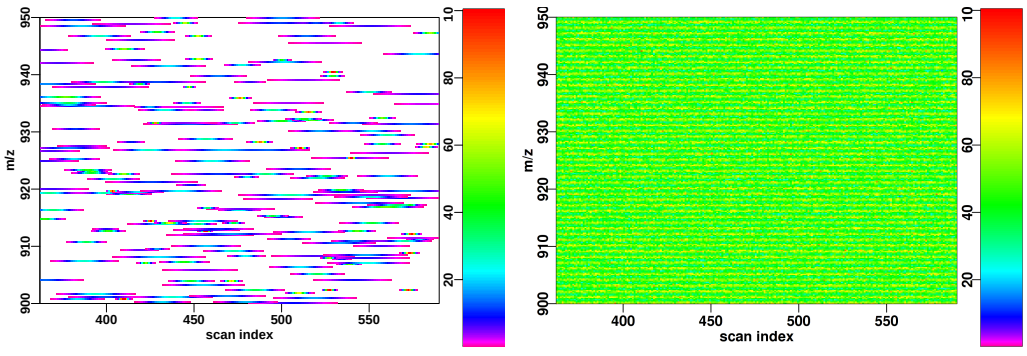

Figure 8.11: Synthetic peptide signals and background noise used to generate a synthetic LC/MS image. The left panel shows the superposition of Gaussian signals. The right panel shows the background noise process.

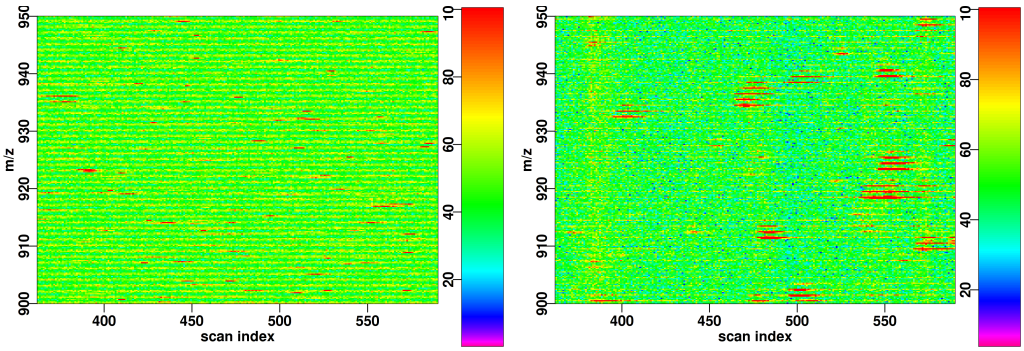

Figure 8.12: The left LC/MS image is the synthetic image. The right LC/MS image shows a real LC/MS image for comparison. In particular, we did not model the isotopic profiles in the synthetic LC/MS image

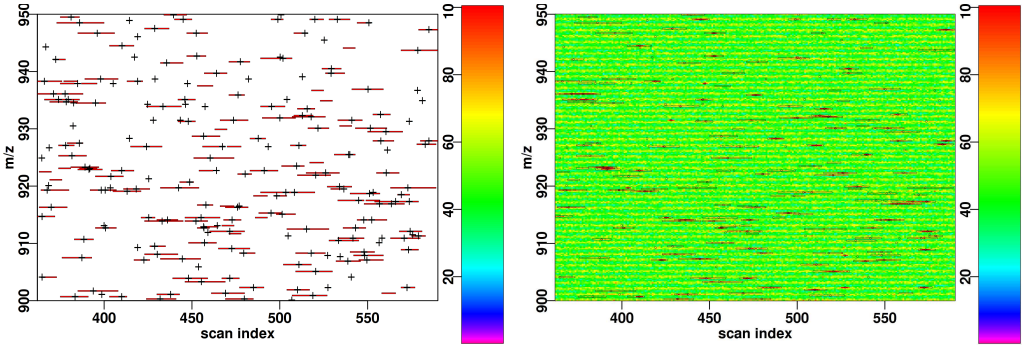

Figure 8.13: Feature detection results on a synthetic LC/MS image. The left panel shows the real peak centroids as black crosses, and the detected pamphlets as red rectangles. The right panel is an alternative representation which makes it possible to view the detection results in their context and evaluate the problem difficulty in a varying background noise.

The two components of the synthetic image in Figure 8.12 were generated in the following way. The peptide component is the sum of 200 peptide signals, with retention time and m/z ratio chosen at random uniformly inside the LC/MS image. The area under the curve A and standard deviation σ are also drawn uniformly inside the plane represented in Figure 8.15, that is to say $A \in [50, 500]$ and $\sigma \in [1, 10]$. The background noise component $\mathcal{N}(t, m)$ was generated such that $\mathcal{N}(t, m)$ is a Poisson random variable with mean $\lambda(t, m) = A\cos(2\pi m) + B$ and A and B are adjusted to match approximately a real LC/MS image as shown in Figure 8.12.

After feature detection, we may evaluate the performance of the feature detection algorithm "visually" with plots similar to Figure 8.13, where we can count the number of true positives and false negatives.

8.6.3 Quantitative evaluation of feature detection

Based on Figure 8.13, we can draw a Venn diagram in Figure 8.14 after counting the detections, false positives, false negatives. This diagram summarizes the results of one experiment, but additional analysis is required to distinguish for example which detections were easy, what types of patterns lead consistently to false positives, etc.

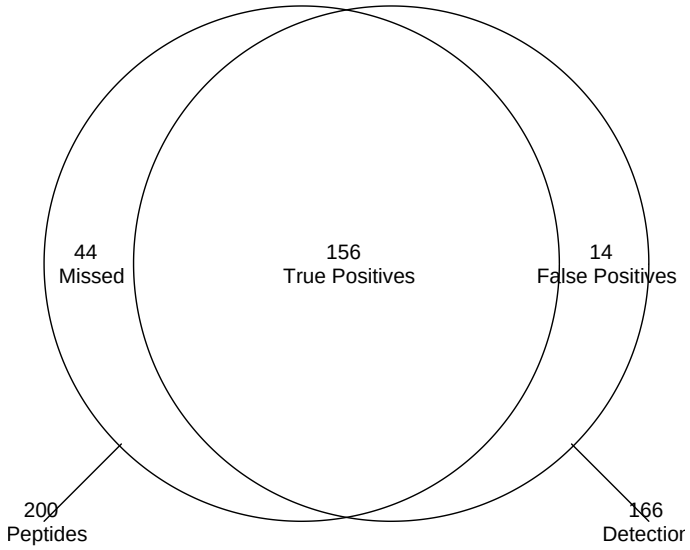

Figure 8.14: A Venn diagram can be used to represent the detection results as well as the missing detections.

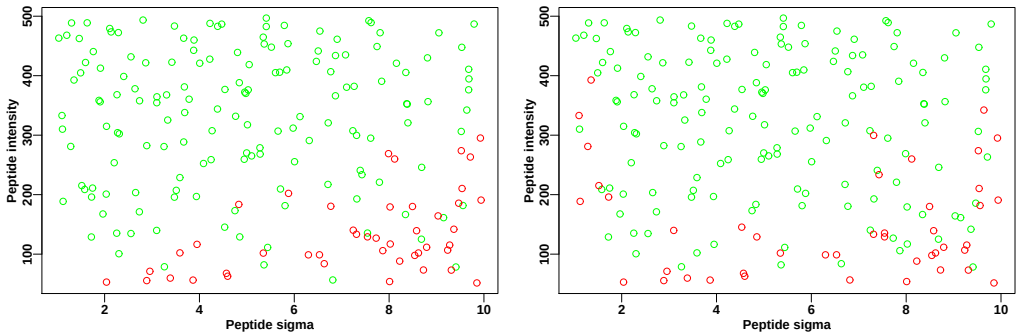

Figure 8.15: Detections results of the quantile M-N rule with $N = 3$ (left) and $N = 7$ (right). The green circles indicate true positives and the red circles indicate false negatives. $N = 7$ is better suited for the detection of larger values of σ, which is coherent with the theoretical results in Section 8.4.2.

Receiver operating characteristic (ROC) curves summarize the performance of a detector at different false positive rates (see [Faw04] for details). However, they do not indicate the performance of a detector as a function of the signal shape. Consequently, we did not compute ROC curves, but rather the plots in Figure 8.15, which provide two types of information:

- they indicate the shapes (A, σ) that the algorithm favors,
- for a given shape (A, σ), we can evaluate the probability of detection by considering the results in the neighborhood of (A, σ)

As show in Figure 8.15, the representation can be used to compare the results of two algorithms. However, fair and meaningful comparisons are obtained only if the false positive rate of the two algorithms is the same.

8.6.4 Notion of detection

There is a major issue with using these quantitative evaluation methods: when do we say that a peak has been detected ? This issue is apparent when looking at the results in Figure 8.13. More precisely, in this figure, some pamphlets clearly are detections because the peak centroid is in the middle of the pamphlet. Others are clearly not detections because they are far from any peak. In between, there are pamphlets for which the corresponding peak is at one end. Should we consider these as true positives or false positives ?

In other areas of image analysis such as edge detection, visual validation is performed on synthetic images or on standard data sets. Quantitative assessment is possible because the features are well located in space, and there is no "middle ground" between true and false positives. Similarly, in image segmentation, the boundaries between regions of the image are well-defined. MS peaks on the contrary have a heavy tail, and the boundary between "peptide" and "no peptide" regions is unclear.

In Figure 8.14 and Figure 8.15, we decided to merge overlapping pamphlets. A resulting region is considered true positive when it contains a peak centroid. We could have been more lenient, and allow some distance between the pamphlet and the peak centroid. This is because the template matching step applied after candidate selection may still retrieve the true position of the centroid, even outside of the candidate region.

We did not compare the M-N rule with other methods because we are not aware of other methods in the field of LC/MS image analysis that detect peaks directly in the LC dimension. Most methods work in the m/z direction (inside a mass spectrum), perform template matching, then assemble the results from neighboring scans. The comparison would not be relevant because our method does not perform template matching.

8.7 Extensions

In this section, we consider some extensions of the feature detection algorithm and of the models for the background noise. In Section 8.7.1, we study the effect of resampling to improve the limit of detection in the M-N rule. In Section 8.7.2, we propose to transform the retention times to exploit a relationship between the retention time of a peptide and its standard deviation.

8.7.1 Resampling the LC/MS image

The M-N rule uses subsampled data, obtained after binning the intensity values in fixed-width bins of size 0.1 Da. This is motivated by:

1. a reduction in computational time and memory requirements for the detection of signals in LC/MS images,
2. the non-uniform sampling rate inside individual mass spectra as well as m/z calibration; resampling or interpolation is necessary to compare successive scans,
3. improvements to the limit of detection obtained by choosing the appropriate bin width as presented in this section.

Non uniform sampling of the m/z axis Depending on the technology of the mass spectrometer, the m/z axis may be uniformly sampled or not (e.g. TOF mass analyzers, see Section 3.3), but in all cases, m/z calibration can prevent the direct comparison of intensity values in neighboring mass spectra. Resampling methods such as binning ensure that the data is sampled at reproducible m/z values. In this section, we ask what are the consequences of resampling, and in particular m/z binning, for the detection of peptide signals.

Problem statement Before detection with the M-N rule, the LC/MS image is binned into fixed-width bins of size u, with u a positive real number. This corresponds to setting the pixel height. We explore the choice of u for the specificity of the M-N rule (false positives) and its effect on the limit of detection (false negatives).

Extension of the noise model We expect tall pixels to contain more background noise intensity than smaller pixels. To quantify this assessment, we model the background noise as a continuous stochastic process. The physics of the experiment suggest a spatial Poisson process. In the following, we present the spatial Poisson process in the context of LC/MS images with a simplified formulation. More details can be found in [Rip04, Kut98, MW04].

Spatial Poisson processes Let $\lambda(t, m)$ be a positive function of retention time t and m/z ratio m that is integrable with respect to the Lebesgue measure. A spatial Poisson process $\mathcal{P}$ with intensity $\lambda(t, m)$ is a random set of points $x_i = (t_i, m_i)$ such that:

1. For all disjoint Borel sets A and B, $\int_A \mathcal{P}$ and $\int_B \mathcal{P}$ are independent random variables.
2. For a Borel set A, $\int_A \mathcal{P}$ is a Poisson random variable with mean parameter $\int_A \lambda(t, m)\mathrm{d}t\mathrm{d}m$.

where $\int_D \mathcal{P} = \#\{x_i \in D\}$ is the number of points inside the domain D.

The spatial Poisson process model justifies some of the assumptions in Section 8.3.2:

- The background noise intensity is the integral $\int_D \mathcal{P}$ over a pixel domain. Consequently figures and simulations were generated based on Poisson distributed background noise.
- Pixel domains are disjoint, and thus the pixel intensities are independent random variables.

The intensity function $\lambda(t, m)$ allow us to derive the background noise intensity as a function of the pixel size. For background noise in LC/MS images, we suppose that the background noise intensity is the same in successive mass spectra (normalization hypothesis) and that the resolution of the mass spectrometer is high enough, so that λ is roughly constant and the Poisson process is uniform locally. The noise intensity in a pixel of height u and retention time range $[t_1, t_2]$ is a Poisson random variable with mean $\lambda u(t_2 - t_1)$.

We approximate the quantiles of the Poisson distribution with

$$q_{\mathcal{N}}(\xi) = \lambda u(t_2 - t_1) + q_0(\xi)\sqrt{\lambda u(t_2 - t_1)}$$

where $q_0(\xi)$ are the quantiles of a normal distribution. Consequently, the mean intensity of the noise is a linear function of the pixel height u. In the following, $t = (t_1 - t_2)$ corresponds to the width of the pixel, i.e. the time needed to acquire a mass spectrum.

The extended noise model verifies the assumptions in 8.4.1. The false positive rate can be controlled with Equation 8.3:

$$H \geq q_{\mathcal{N},1-\alpha^{1/N}}$$

The height u of the pixel has no consequence on the false positive rate of the M-N rule.

Intensity model Let A denote the total area of the peptide signal in the 2D image. Then the (continuous) peptide signal a separable product of two Gaussian functions:

$$I(t, m) = A\Gamma_1(t)\Gamma_2(m)$$

In a m/z bin of width u, we consider that the peptide signal intensity is

$$I(t) = \left(A \int_{-u/2}^{u/2} \Gamma_2(m) \mathrm{d}m \right) \Gamma_1(t)$$

which corresponds to the model presented in (8.3) with area under the curve $A \int_{-u/2}^{u/2} \Gamma_2(m) \mathrm{d}m$.

The signal-to-noise ratio defined in Equation (8.5) becomes:

$$\frac{A \int_{-u/2}^{u/2} \Gamma_2(m) \mathrm{d}m}{q_{\mathcal{N},1-\alpha^{1/N}} - q_{\mathcal{N},1-(1-\beta)^{1/N}}} \geq \sqrt{2\pi\sigma^2} \exp\left\{ \frac{N^2/4}{2\sigma^2} \right\}$$

which we can write in terms of limit of detection:

$$A \geq \frac{q_{\mathcal{N},1-\alpha^{1/N}} - q_{\mathcal{N},1-(1-\beta)^{1/N}}}{\left(\int_{-u/2}^{u/2} \Gamma_2(m) \mathrm{d}m \right) \Gamma_\sigma(N/2)}$$

Optimum of the pixel height Fix (A, σ) a peptide signal, and $\Gamma_2(m)$ a Gaussian function with standard deviation 1. Fix the parameters of the M-N rule detector. The limit of detection is a function of u:

$$B : u \mapsto \frac{\sqrt{\lambda u t} \left(q_0(1-\epsilon^{1/N}) - q_0(1-(1-\beta)^{1/N}) \right)}{\int_{-u/2}^{u/2} \Gamma_2(m) \mathrm{d}m \Gamma_\sigma(N/2)}$$

In particular, the optimal choice for u is independent of the parameters of the M-N rule detector.

Limits Before optimizing, we study the limits when u is small and u tends to $+\infty$.
Let $C = \dfrac{\left(q_0(1-\epsilon^{1/l}) - q_0(1-(1-\beta)^{1/N}) \right)}{\Gamma_\sigma(N/2)}$.
When $u \to \infty$, $B(u)$ is equivalent to $C\sqrt{\lambda u t}$ because $\int_{-u/2}^{u/2} \Gamma_2(m) \mathrm{d}m \to 1$. Consequently, the detection limit tends to infinity, which means that the signal is lost in the noise.
When $u \to 0$, $B(u)$ is equivalent to $\dfrac{C\sqrt{\lambda u t}}{\Gamma_2(0) u} = \dfrac{C\sqrt{\lambda t}}{\Gamma_2(0)\sqrt{u}}$. When u is too small, the signal is also lost in the noise.

Approximation Let us assume that the 2D peptide shape in the m/z direction is not Gaussian, but uniform on $[-a/2, a/2]$. In other words, we replace $\Gamma_2(m)$ by $\frac{1}{a} \mathbb{1}_{[-a/2,a/2]}$. Then

$$B(u) = \begin{cases} \dfrac{\sqrt{\lambda t}\, C\, a}{\sqrt{u}} & \text{if } u \leq a \\ \sqrt{\lambda u t}\, C & \text{if } u \geq a \end{cases}$$

The optimum value for u is $u = a$. This choice corresponds to $u = \sqrt{12}\,\sigma_2 \sim 3.46\,\sigma_2$ where σ_2 is the standard deviation of Γ_2. Similarly, more precise estimates can be obtained with higher order approximations of the Gaussian distribution, but we did not make the computations.

By numerical minimization of B, we obtain the following estimate $u = 2.80\sigma_2$. This choice is independent of the other parameters: peptide shape (A, σ), detector parameters (H, N) or significance levels α, β. Consequently, the pixel height can be chosen a priori, based on the resolution of the mass spectrometer.

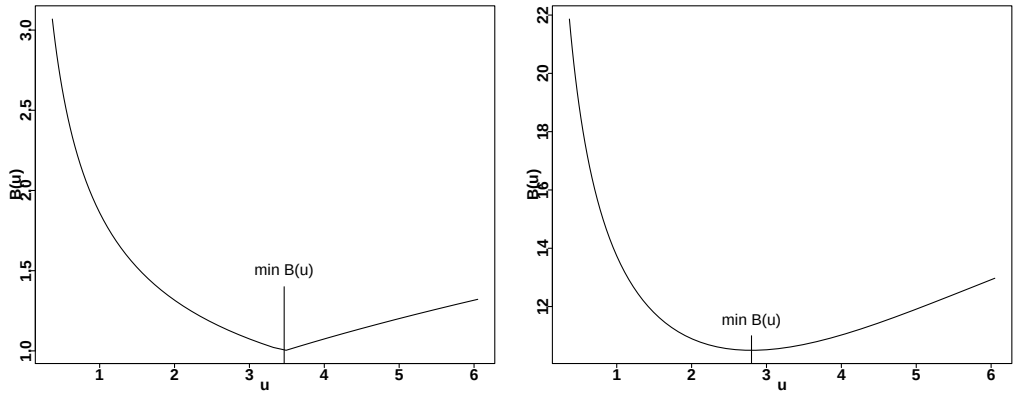

Figure 8.16: Lower limit of detection as a function of u. The left panel corresponds to an approximation of Γ_2 with a uniform distribution. The right panel uses the exact formula.

In Figure 8.17, we compare the signal intensity of a Gaussian signal (in black) with samples of a Gaussian noise distribution (in red) with mean λut and standard deviation $\sqrt{\lambda ut}$ for several values of u. For $u = 8$ on the right panel, detection of the Gaussian signal is difficult because the overall noise intensity is too high. For $u = 0.3$ in the left panel, the noise level is not very high as compared to the signal intensity, but the variability of the noise makes detection difficult. The optimum value $u = 2.80$ is shown in the middle panel.

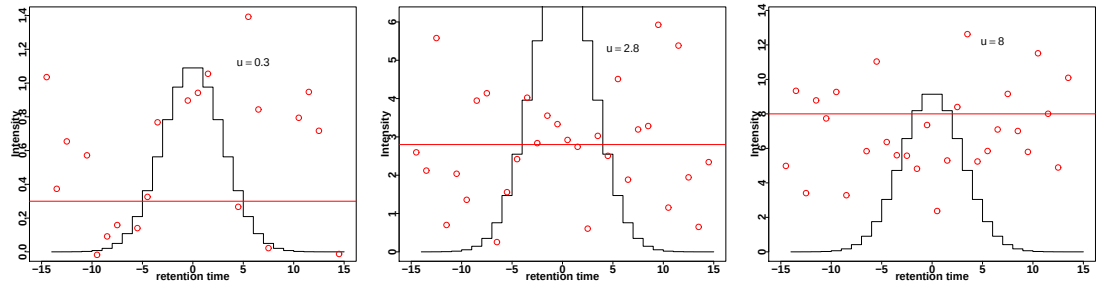

Figure 8.17: Comparison of the intensity of a sampled Gaussian signal (black) and the intensity of Gaussian noise (red) for several values of the pixel height u. The red line indicates the noise level (mean noise intensity).

Conclusion Prior to feature detection with the M-N rule, the LC/MS image is resampled into pixels. The width of the pixels is set to one mass spectrum, but the height of the pixel can be chosen. We show that detection can be improved by selecting the inverse of the resolution of the mass spectrometer as the pixel height.

Poisson point processes are probably good models for centroided data (see Figure 6.6 on page 6.6). However, the behavior of centroiding algorithms is unspecified and implemented in manufacturer software or even hardware. It is not clear what type of point process is generated by centroiding, for example standard point processes or compound processes.

8.7.2 Compensation for varying peak shape

As described in Section 8.3, there is a relationship between the retention time μ_k of a peptide signal and its standard deviation σ_k :

$$\sigma_k = \mu_k/\sqrt{P}$$

where P is the plate number of the chromatography column. In this section, we investigate a method to simplify feature detection based on this relationship. Note that this is only valid for isocratic mode [2] and as such was not used in the submitted manuscript.

Problem statement Given a function $\sigma(t)$, compute a transformation f of the retention time axis that preserves area under the curve, and study its effects on feature detection with the M-N rule.

Retention time transformation We look for a function f such that $\sigma(f(t)) = 1$, so the natural choice is σ^{-1}. In fact, this is not a solution to the problem because σ^{-1} transforms local lengths, but f should transform the retention time coordinates. Instead, consider that $\sigma(t)$ is the width of a Gaussian signal at retention time t:

$$\sigma = \int_{-\sigma/2}^{\sigma/2} 1 \mathrm{d}x$$

Gaussian signals have the same width if

$$\sigma = \int_{-1/2}^{1/2} f'(y) \mathrm{d}y$$

where $f'(y) = \dfrac{1}{\sigma(y)}$ is an approximate solution to the problem.

To also preserve the area under the curve, we consider the following transformation:

$$\mathcal{I}(t) \mapsto \frac{\mathcal{I}(f(x))}{f'(x)}$$

In the case of $\sigma(t) = t/\sqrt{P}$, we obtain $f : t \mapsto \log t$

This logarithmic transform has been proposed in [THN+04] without justification, and applied in the m/z direction to correct variations in the peak shape. In the current context, we apply the transformation in the retention time direction, and the log transform is justified by the relationship between the retention time of a peptide and its standard deviation.

Figure 8.18 shows the effect of warping the retention time on a signal made of a superposition of Gaussian peaks with $\sigma = \mu$ and with the same area under the curve. After transformation of the retention times with $f : t \mapsto \log t$, the elution profiles of each signal have the same standard deviation.

Note that plotting the signal $\dfrac{\mathcal{I}(f(x))}{f'(x)}$ does not perform resampling per se. To do that, we need to write a function that adds the intensity values while preserving area. The step function in Figure 8.18 is the result of true resampling.

Noise transformation The proposed warping of the retention times modifies the noise intensity as a function of retention time. Let $\lambda(t)$ denote the mean noise intensity, then after transformation, it becomes $\sigma^2(t)\lambda(t)$. For $\sigma = t/\sqrt{P}$, noise intensity increases as shown in Figure 8.19.

[2] Gradient elution (described in Chapter 2), is a method in RPLC chromatography that modifies the solvent composition during the course of the experiment. In isocratic mode, the solvent composition is not modified.

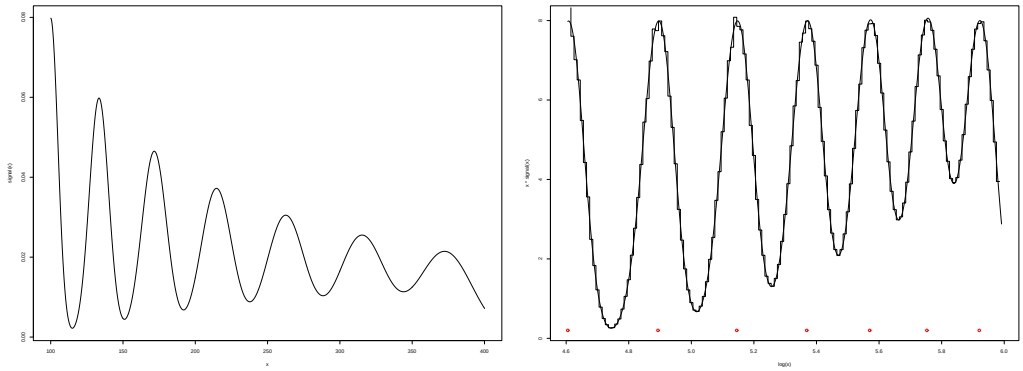

Figure 8.18: Transformation of the retention time axis. A superposition of Gaussian functions (shown on the left panel) is transformed into a series of identical signals (right panel). The ragged line is the transformation of the signal after pixellisation. Red dots indicate the new centroids.

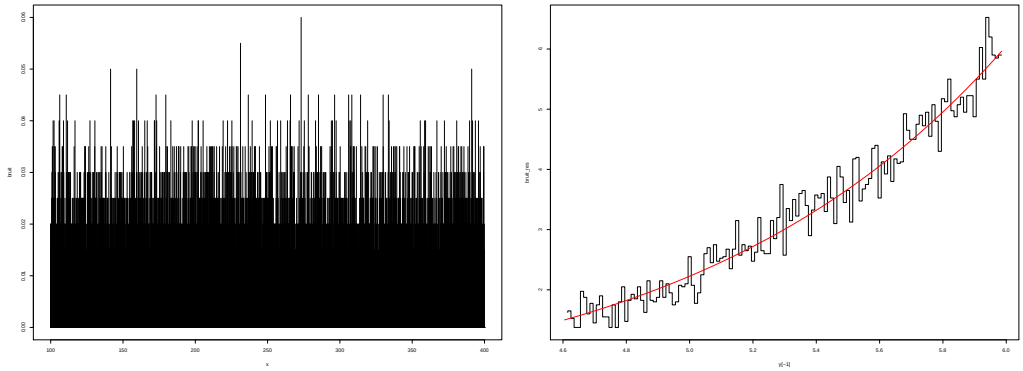

Figure 8.19: Transformation of retention times applied to random noise. The left panel shows Poisson random variables. The right panel shows the results of resampling the intensity function. The red line corresponds to the predicted mean intensity of the noise after resampling.

Modification of the threshold As the background noise is not stationary in the m/z bin, we have to adapt the threshold H used in feature detection with the M-N rule. Figure 8.20 shows the adapted threshold for feature detection.

Conclusion Based on the relationship between retention time and standard deviation, we propose a transformation of the retention times to reduce the variability of the signal shapes. We show that after transformation, the width of the peptide signals are constant. Therefore, we can select a universal value for the parameter N in the M-N rule. For peak picking, the same template can be matched to all signals in the LC/MS image.

The logarithmic transform is limited to isocratic mode, because other gradients modify the retention of the peptides on the RPLC column. It is unknown if a relation between retention time and standard deviation exists in LC/MS images obtained in gradient elution, or if it can be used to improve signal detection as proposed in this section.

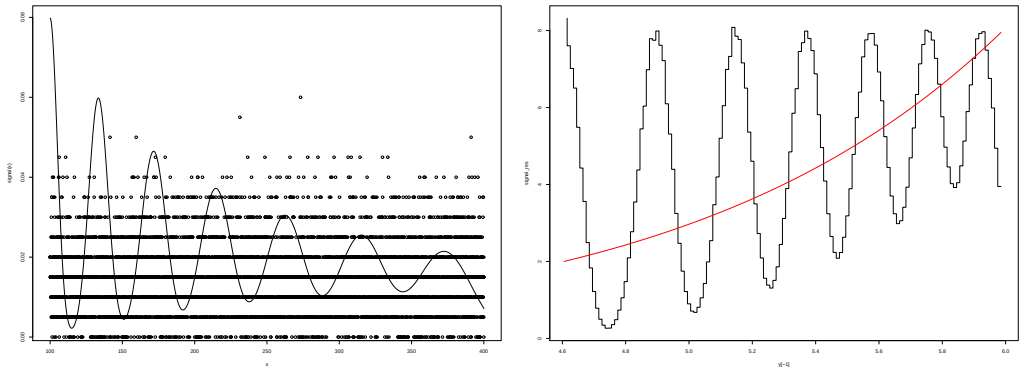

Figure 8.20: Transformation of retention times for feature detection. The left panel shows the relative intensities of noise and signal before transformation. The right panel shows the threshold and signal after transformation.

8.8 Related problems

The M-N rule consists in applying multiple, local statistical tests to different locations in the LC/MS image to detect peptide signals. More precisely, it corresponds to testing if the local minimum of the pixel intensity is above a threshold. Many other local descriptors of the intensity can be suggested instead of the minimum: local maximum, mean, median, quartile, variance, along with their robust versions.

Background material on the topics discussed in this section can be found in the introductory chapters of [dFCJ01]. A list of measures of location (mean) and scale (variance) can be found in [MMY$^+$06] from the point of view of statistics. In the image analysis literature, the computation of local characteristics is often called *filtering*, and a list of classical approaches (linear filters, non-linear filters, adaptive filters, etc.) can be found in [CPB95].

8.8.1 Rank filters

The M-N rule is naturally viewed as thresholding first, and looking for correlation in neighboring mass spectra afterwards. Its essence is in the search for groups of contiguous pixels which indicate local correlation. The detection results can also be computed by scanning the horizontal line for successions of pixels of at least length N.

In the case of the M-N rule, thresholding and filtering can be reversed to fit in the framework of rank filters, i.e. we can first compute a rank filter (the local minimum) then apply thresholding. However, the M-N rule does not really correspond to this context because rank filters are used for denoising rather than detection and "are mainly useful for removing impulse noise" [CPB95]. Consequently rank filters are usually based on the median rather than the extremes.

The strict decision rule $T \geq N$ can be replaced with $T \geq N - k$ which authorizes up to k pixels under the threshold. In the context of rank filters, this corresponds to using the k-th lowest value instead of the minimum. More generally, this corresponds to increasing the search space in the feature detection algorithm.

Increasing the search space requires a tighter control of the false positive rate per hypothesis. For example, [DMM08] proposes to find maximal significant intervals in a signal [3], i.e. sets of contiguous pixels that with probability less than α but that cannot be extended to the left or to the right. The search space contains w^2 pamphlets, that is to say all possible segments of all length in a horizontal line of w pixels, or equivalently w mass spectra in the LC/MS image. In comparison, the M-N rule considers only on the order of w different pamphlets.

8.8.2 Mathematical morphology

It is more appropriate to view the M-N rule in the context of mathematical morphology (see [CS05, Ser82, Ser88]) rather than rank filtering. Indeed, the M-N rule corresponds to an erosion filter in one dimension. Also, the properties developed in this chapter are similar to morphological filters, which select shapes and other features in the image depending on their scale.

Mathematical morphology suggests other possible choices to replace the erosion operator, such as dilation (local maximum, often used in feature detection), opening and closing. The connections with mathematical morphology still remain to be fully explored, as well as the connections with variational approaches (see [CS05] for example).

8.8.3 Linear filters

Instead of rank filters which are non-linear, we could have used the theory of linear filters. This theory is appealing because we can compute the filter with optimal signal to noise ratio, with the hypothesis that the signal is known in advance and that the noise is white Gaussian noise. This filter is a symmetrised version of the input signal (see [CPB95] for the full derivation). The linear theory has been particularly successful in the case of Gaussian white noise, and many results can be found e.g. in [VT01].

In the case of LC/MS images, we chose the M-N rule to build upon an existing method, and also because it is not tied to a specific noise distribution. In LC/MS data, the background noise is rather Poisson distributed and varies in different locations of the image. Moreover, the linear filter theory is focused on a known signal shape. Developments similar to those in Section 8.4.2 are necessary to evaluate linear filters for the detection of shapes with varying characteristics.

8.8.4 Replicates

Feature detection with the M-N rule does not take into account the replicates of a protein signal (isotopes and charge states) for detection. This requires combining several detectors which is an interesting question in multiple testing.

The combination of several feature detectors is complicated by the fact that the replicates are not identical. For instance:

- signals from isotopes and other charge states usually have lower intensity than the monoisotopic peak,
- the distribution of intensity between isotopic peaks is dependent on the chemical composition of the molecule.

[3] For histogram segmentation

8.9 Conclusion

The M-N rule is a feature detection algorithm that can effectively recall low-intensity peptide signals in LC/MS images. In this chapter, we extend the original formulation and provide a precise account of the effect of the new parameters H and N on the detection results. N controls the standard deviation of the elution profiles that can be detected reliably, while H controls the selectivity of the algorithm. We provide guidelines for choosing N for a given LC/MS image, and compute H from the local noise distribution. The resulting Quantile M-N rule is guaranteed to yield a false positive rate bounded by a user-defined parameter α.

In the context of LC/MS images, we develop the theory of a contrario detection in the new direction of sensitivity, with a geometric model of the shapes to be detected. In contrast to other detection problems like edge detection or shape recognition, the set of shapes is parametrized by only the area under the curve A and the standard deviation σ. Additionally, we have a well-defined goal, which is to optimize the detection of low-intensity shapes.

Due to the width of the features considered in LC/MS images, it is difficult to locate the peaks precisely and define what a detection is. Several evaluation procedures are discussed in Section 8.6. We propose to use plots in feature space similar to Figure 8.15 to compare different methods, but in practice, these are difficult to draw at the same false positive rate.

The proposed extensions provide answers to two additional questions. We first address the question of the choice of the pixel height for resampling the LC/MS image in case of irregular sampling. In that case, resampling is necessary but we also show that there is an optimal choice for the pixel height that optimizes the detection of peptide signals. We then address the formula 8.1. If this relationship between retention time and the standard deviation of elution profile is known a priori, then it suffices to select different values of N. Instead we show that the LC/MS image can be resampled to obtain the same standard deviations for all peaks, and use only one set of parameters. However, it is not clear how this improves detection because the noise intensity in the resampled image is no longer stationary. In Section 8.8, we put the M-N rule into the wider context of filtering theory and mathematical morphology. This allows us to sketch options for extending the M-N rule.

Feature detection with the M-N rule does not provide precise values for the signal retention time and m/z ratio. It does not tackle the deconvolution of overlapping features, nor does it take isotope patterns into account. When using the M-N rule, these tasks need to be handled by another algorithm with greater attention to individual pixel intensities in the pamphlet. However, the M-N rule can be used as a filtering step prior to accurate peak picking to provide a statistical control of the false positive rate.

Appendix

Sensitivity of the M-N Rule

In this section, we justify the computation of the sensitivity of the extended M-N rule. In a horizontal line, the recorded intensity is modelled as:

$$\mathcal{I}(t) = \mathcal{N}(t) + A\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{t^2}{2\sigma^2}\right\}$$

where $\mathcal{N}(t)$ is the background noise and the signal is Gaussian, with area under the curve A, standard deviation σ and is centred in the pamphlet.

Given a false positive rate α and the M-N rule parameters $H = q_{\mathcal{N},1-\alpha^{1/N}}$ and N, we compute the probability that a Gaussian signal with area under curve A and standard deviation σ is detected by the M-N rule:

$$\begin{aligned}
\mathbb{P}\left[\mathcal{I}_i > H, \forall i \in \{1,\dots,N\}\right] &= \prod_{i\in\{1,\dots,N\}} \mathbb{P}\left[\mathcal{I}_i > H\right] \\
&= \prod_{i\in\{1,\dots,N\}} \mathbb{P}\left[\mathcal{N} + \frac{A}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{t_i^2}{2\sigma^2}\right\} > H\right] \\
&\geq \prod_{i\in\{1,\dots,N\}} \mathbb{P}\left[\mathcal{N} + \frac{A}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{N^2/4}{2\sigma^2}\right\} > H\right] \\
&= \left(\mathbb{P}\left[\mathcal{N} + \frac{A}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{N^2/4}{2\sigma^2}\right\} > H\right]\right)^N
\end{aligned}$$

The lower bound is obtained because the Gaussian function $\Gamma(t)$ is above the value $\Gamma(N/2)$ in the range $t \in [-N/2; N/2]$.

Given the false negative rate β, we solve the following inequality:

$$\begin{aligned}
&\left(\mathbb{P}\left[\mathcal{N} + \frac{A}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{N^2/4}{2\sigma^2}\right\} > H\right]\right)^N \geq 1-\beta \\
\Leftrightarrow\quad &\mathbb{P}\left[\mathcal{N} + \frac{A}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{N^2/4}{2\sigma^2}\right\} > H\right] \geq (1-\beta)^{1/N} \\
\Leftrightarrow\quad &\mathbb{P}\left[\mathcal{N} + \frac{A}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{N^2/4}{2\sigma^2}\right\} < H\right] \leq 1-(1-\beta)^{1/N} \\
\Leftrightarrow\quad &H - \frac{A}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{N^2/4}{2\sigma^2}\right\} \leq q_{\mathcal{N},1-(1-\beta)^{1/N}}
\end{aligned}$$

In conclusion we obtain:

$$\begin{aligned}
&H - q_{\mathcal{N},1-(1-\beta)^{1/N}} \leq \frac{A}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{N^2/4}{2\sigma^2}\right\} \\
\Rightarrow\quad &\mathbb{P}\left[\mathcal{I}_i > H, \forall i \in \{1,\dots,N\}\right] \geq 1-\beta
\end{aligned}$$

Chapter 9

Conclusions and Perspectives

9.1 Conclusions

This manuscript studies signal distortions and methods to correct them for LC/MS images acquired with a mass spectrometer coupled to a liquid chromatography separation. On this instrument platform, systematic sources of variation in the sample preparation, the liquid chromatography as well as the mass spectrometer can hinder the interpretation of the data for proteomics applications such as identification of proteins, quantification and biomarker discovery.

In this thesis, we provide a review of the technologies for LC/MS analysis of biological samples and their related signal distortions. Although we have focused on Q-TOF instruments, most distortions are in fact common to all LC/MS platforms, albeit with different experimental justification. We review the methods for correcting the observed retention times in Chapter 6 and develop original approaches for correcting intensity values and detecting peptide signals based on the background chemical noise.

Without a clear understanding of the statistical properties of the background noise in the literature, we have studied empirical approaches for intensity normalization and baseline estimation in Chapter 6 as well as model-based approaches in Chapter 7. These methods build on the hypothesis that noise in LC/MS images is generated by real molecules, and that it is affected by the normalization factor. Their potential to improve on current methods lies in the fact that peptide signals are separated and thus isolated in LC/MS images; most of the data set is actually background noise.

Peptide signals can therefore be detected as significant deviations from the background noise. In Chapter 8, we study a feature detection algorithm that applies local statistical tests to the pixel intensities in the LC/MS image. Inside this mathematical framework, we show how to control the false positive rate of detection and how to optimize the parameters based on a signal-to-noise ratio adapted to the algorithm.

For the proposed methods, we have tried to provide mathematical results that are relevant for adoption in the field of LC/MS-based proteomics. For instance, our feature detection scheme extends an already available method, but studies carefully the artifacts (false positives) and the limit of detection (false negatives) of the method. The algorithm is quick, easy to implement and also easy to re-use in different contexts with the same statistical properties.

9.2 Perspectives

Although a lot of research has been done on signal preprocessing for LC/MS images, some fundamental challenges remain. First, multimodal comparison is a major issue. Most methods are designed to erase systematic variations between LC/MS images that are thought of as ideally identical. With biological samples from different contexts (patients and control cases for instance), signal processing methods may eliminate legitimate differences and prevent the discovery of biomarkers. Such problems include:

- forced alignment of signals from unrelated peptides during the correction of retention times,
- inclusion of low intensity signals in the baseline component,
- normalization of the intensities of samples with different protein concentrations. For instance, elevated protein concentrations in tumor cells is a feature and not artifact.

From a practical point of view, most of the approaches presented in this document are faced with very large datasets and steep computer hardware requirements in terms of processor speed and random access memory. When our thesis began, back in late 2005, we could barely compare a full resolution LC/MS image before and after signal processing, as two images would not fit inside the two gigabytes of available RAM. Since then, the price of memory chips has dwindled and we purchased an additional two gigabytes of RAM. We also decided to subsample the data as most workstations would not meet the requirements (most still don't).

The mathematical extensions to the work in this thesis have been largely discussed in the individual chapters and will not be reproduced here. However, we would like to insist on two remarks. First, the standard analysis pipeline as indicated in the introduction in Figure 1.3, is not set in stone. For example, biomarker discovery approaches usually identify peptide signals before quantification, but some approaches test for differences between LC/MS images even before feature detection. There is also potential for doing protein identification based on both MS/MS and retention time. Second, statistical analysis of LC/MS data should take into account the preprocessing steps to avoid propagating biases and errors. For instance, our feature detection scheme is independent of baseline substraction and normalization, and can be extended to take smoothing into account.

Chapter 10

Bibliography

[AAC+05] R. Aebersold, L. Anderson, R. Caprioli, B. Druker, L. Hartwell, and R. Smith. Perspective: a program to improve protein biomarker discovery for cancer. *J Proteome Res*, 4(4):1104–1109, 2005.

[ABF99] A. Antoniadis, J. Berruyer, and A. Filhol. Estimation semi-paramétrique dans les familles doublement poissonniennes et application aux spectres de diffraction. *Revue de Statistique Appliquée*, 47(3):57–80, 1999.

[ACPS06] D.B. Allison, X. Cui, G.P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7(1):55–65, Jan 2006.

[ACvG+06] A.H.P. America, J.H.G. Cordewener, M.H.A. van Geffen, A. Lommen, J.P.C. Vissers, R.J. Bino, and R.D. Hall. Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional lc-ms. *Proteomics*, 6(2):641–653, Jan 2006.

[AM03] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, Mar 2003.

[ARC+03] V.P. Andreev, T. Rejtar, H.S. Chen, E.V. Moskovets, A.R. Ivanov, and B.L. Karger. A Universal Denoising and Peak Picking Algorithm for LC- MS Based on Matched Filtration in the Chromatographic Time Domain. *Anal. Chem*, 75(22):6314–6326, 2003.

[ARL+04] M. Anderle, S. Roy, H. Lin, C. Becker, and K. Joho. Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics*, 20(18):3575–3582, 2004.

[ASNN05] T. Aittokallio, J. Salmi, T.A. Nyman, and O.S. Nevalainen. Geometrical distortions in two-dimensional gels: applicable correction methods. *J Chromatogr B Analyt Technol Biomed Life Sci*, 815(1-2):25–37, Feb 2005.

[BAVL06] G. Bhanot, G. Alexe, B. Venkataraghavan, and A.J. Levine. A robust meta-classification strategy for cancer detection from ms data. *Proteomics*, 6(2):592–604, Jan 2006.

[BCF+06] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C.W. Lin, et al. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*, 22(15):1902, 2006.

[BCM05] K.A. Baggerly, K.R. Coombes, and J.S. Morris. Bias, randomization, and ovarian proteomic data: A reply to "producers and consumers". *Cancer Informatics*, 1(1):9–14, 2005.

[BDA+07] V. Brun, A. Dupuis, A. Adrait, M. Marcellin, D. Thomas, et al. Isotope-labeled protein standards: toward absolute quantitative proteomics. *Molecular & Cellular Proteomics*, 6(12):2139, 2007.

[BDMM02] D. Bylund, R. Danielsson, G. Malmquist, and K.E. Markides. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for parafac modelling of liquid chromatography-mass spectrometry data. *J Chromatogr A*, 961(2):237–244, Jul 2002.

[BH96] T.W. Burgoyne and G.M. Hieftje. An introduction to ion optics for the mass spectrograph. *Mass Spectrometry Reviews*, 15(4), 1996.

[BHWW00] E.J. Breen, F.G. Hopwood, K.L. Williams, and M.R. Wilkins. Automatic poisson peak harvesting for high throughput protein identification. *Electrophoresis*, 21(11), 2000.

[BM08] S. Bocker and V. Makinen. Combinatorial approaches for mass spectra recalibration. *IEEE ACM Transactions on Computational Biology and Bioinformatics*, 5(1):91, 2008.

[BMEC05] K.A. Baggerly, J.S. Morris, S.R. Edmonson, and K.R. Coombes. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *J Natl Cancer Inst*, 97(4):307–309, Feb 2005.

[BMW03] KA Baggerly, JS Morris, and J. Wang. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionizationtime of flight proteomics spectra from serum samples. *Proteomics*, 3:1667–1672, 2003.

[BPSW70] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math Stat.*, 41:164–171, 1970.

[Bro96] R. Bro. Multiway calibration. multilinear pls. *Journal of Chemometrics*, 10(1), 1996.

[Bro97] R. Bro. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, 1997.

[BS05] B. Bogdanov and R.D. Smith. Proteomics by fticr mass spectrometry: top down and bottom up. *Mass Spectrom Rev*, 24(2):168–200, 2005.

[BSS+07] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, 389(4):1017–1031, 2007.

[BWB03] T.A. Baker, J.D. Watson, and S.P. Bell. *Molecular Biology of the Gene*. Benjamin-Cummings Publishing Company, 2003.

[CBFC06] D. Clifford, M. Buckley, K.Y.C. Fung, and L. Cosgrove. Interactive feature finding in liquid chromatography mass spectrometry data. *Journal of Proteome Research*, 5(11):3179–3185, 2006.

[CE01] N. B. Cech and C. G. Enke. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrom Rev*, 20:362–387, 2001.

[CFP03] M.T. Chu, R.E. Funderlic, and R.J. Plemmons. Structured low rank approximation. *Linear Algebra and its Applications*, 366:157–172, 2003.

[Cha03] L. Charles. Flow injection of the lock mass standard for accurate mass measurement in electrospray ionization time-of-flight mass spectrometry coupled with liquid chromatography. *Rapid Communications in Mass Spectrometry*, 17(13), 2003.

[CJHM07] M.C. Codrea, C.R. Jimenez, J. Heringa, and E. Marchiori. Tools for computational processing of lc-ms datasets: a user's perspective. *Comput Methods Programs Biomed*, 86(3):281–290, Jun 2007.

[Cle79] W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

[CLT01] I.V. Chernushevich, A.V. Loboda, and B.A. Thomson. An introduction to quadrupole-time-of-flight mass spectrometry. *Journal of mass spectrometry*, 36(8):849–865, 2001.

[CM05] Daniel Chamrad and Helmut E Meyer. Valid data from large-scale proteomics studies. *Nat Methods*, 2(9):647–648, Sep 2005.

[CPB95] J.P. Cocquerez, S. Philipp, and P. Bolon. *Analyse d'images: filtrage et segmentation*. Masson, 1995.

[CR03] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Comput. Vis. Image Underst.*, 89(2-3):114–141, 2003.

[CS05] T.F. Chan and J. Shen. *Image Processing and Analysis: variational, PDE, wavelet, and stochastic methods*. Society for Industrial Mathematics, 2005.

[CTM+05] K.R. Coombes, S. Tsavachidis, J.S. Morris, K.A. Baggerly, M.C. Hung, and H.M. Kuerer. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5(16), 2005.

[Cur99] L.A. Currie. Detection and quantification limits: origins and historical overview. *Analytica Chimica Acta*, 391(2):127–134, 1999.

[DA06] B. Domon and R. Aebersold. Mass spectrometry and protein analysis, 2006.

[DDY03] A.W. Dowsey, M.J. Dunn, and G.-Z. Yang. The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics*, 3(8):1567–1596, Aug 2003.

[dFCJ01] Luciano da Fontoura Costa and Roberto Marcondes Cesar Jr. *Shape analysis and classification: theory and practice*. CRC press, 2001.

[DHM+06] N.J. Dovichi, S. Hu, D. Michels, D. Mao, and A. Dambrowitz. *Proteomics for Biological Discovery*, chapter 12 Single Cell Proteomics, pages 225–246. Wiley-Liss, 2006.

[Dia04a] E.P. Diamandis. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *J Natl Cancer Inst*, 96(5):353–356, Mar 2004.

[Dia04b] E.P. Diamandis. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol Cell Proteomics*, 3(4):367–378, Apr 2004.

[DKL06] P. Du, W.A. Kibbe, and S.M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059, 2006.

[DMM00] A. Desolneux, L. Moisan, and J.M. Morel. Meaningful alignments. *International Journal of Computer Vision*, 40(1):7–23, 2000.

[DMM01] A. Desolneux, L. Moisan, and J.M. Morel. Edge detection by Helmholtz principle. *Journal of Mathematical Imaging and Vision*, 14(3):271–284, 2001.

[DMM08] A. Desolneux, L. Moisan, and JM Morel. *From Gestalt Theory to Image Analysis, a Probabilistic Approach, volume 34 of Interdisciplinary Applied Mathematics*. Springer, 2008.

[DMW04] D. Day, A.H. Millar, and J. Whelan. *Plant mitochondria: from genome to function*. Kluwer Academic Publishers, 2004.

[DSPA07] P. Du, R. Sudha, M.B. Prystowsky, and R.H. Angeletti. Data reduction of isotope-resolved LC-MS spectra. *Bioinformatics*, 23(11):1394, 2007.

[EF07] J. Eriksson and D. Fenyo. Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs. *Nat Biotechnol*, 25(6):651–655, Jun 2007.

[EFLA07] L. Elo, S. Filen, R. Lahesmaa, and T. Aittokallio. Reproducibility-optimized test statistic for ranking genes in microarray studies. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2007.

[Efr86] B. Efron. Double exponential families and their use in generalized linear regression. *J. AM. STAT. ASSOC.*, 81(395):709–721, 1986.

[EG07] J.E. Elias and S.P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4(3):207–214, Mar 2007.

[EGW87] K. Esbensen, P. Geladi, and S. Wold. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2:37–52, 1987.

[EHFG05] J.E. Elias, W. Haas, B.K. Faherty, and S.P. Gygi. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods*, 2(9):667–675, Sep 2005.

[Eil03] P.H.C. Eilers. A perfect smoother. *Anal Chem*, 75(14):3631–3636, Jul 2003.

[Eil04] P.H.C. Eilers. Parametric time warping. *Anal Chem*, 76(2):404–411, Jan 2004.

[EMY94] Jimmy K. Eng, Ashley L. McCormack, and John R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976 – 989, 1994.

[EY36] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[Faw04] T. Fawcett. ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, 31, 2004.

[FC80] F. N. Fritsch and R. E. Carlson. Monotone piecewise cubic interpolation. *SIAM J. Numer. Anal.*, 17(2):238–246, April 1980.

[FGR+06] B. Fischer, J. Grossmann, V. Roth, W. Gruissem, S. Baginsky, and J.M. Buhmann. Semi-supervised lc/ms alignment for differential proteomics. *Bioinformatics*, 22(14):e132–e140, Jul 2006.

[Fit97] G.M. Fitzmaurice. Model Selection with Overdispersed Data. *The Statistician*, 46(1):81–91, 1997.

[FRS+07] E.J. Foss, D. Radulovic, S.A. Shaffer, D.M. Ruderfer, A. Bedalov, D.R. Goodlett, and L. Kruglyak. Genetic basis of proteome variation in yeast. *Nat Genet*, 39(11):1369–1375, October 2007.

[FT06] S. Friedland and A. Torokhti. Generalized rank-constrained matrix approximations. *Arxiv preprint math.OC/0603674*, 2006.

[Gas97] S.J. Gaskell. Electrospray: principles and practice. *Journal of Mass Spectrometry*, 32(7), 1997.

[GD90] A. E. Gelfand and S. R. Dalal. A note on overdispersed exponential families. *Biometrika*, 77(1):55–64, 1990.

[Gha90] M. Ghanbari. The cross-search algorithm for motion estimation [image coding]. *IEEE Transactions on Communications*, 38(7):950–953, 1990.

[GMG+99] R. Gras, M. Muller, E. Gasteiger, S. Gay, P.A. Binz, W. Bienvenut, C. Hoogland, J.C. Sanchez, A. Bairoch, D.F. Hochstrasser, et al. Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis*, 20(18), 1999.

[GRB+07] T. Guina, D. Radulovic, A. J. Bahrami, D. L. Bolton, L. Rohmer, K. A. Jones-Isaac, J. Chen, L. A. Gallagher, B. Gallis, S. Ryu, G. K. Taylor, M. J. Brittnacher, C. Manoil, and D. R. Goodlett. MglA regulates Francisella tularensis subsp. novicida (Francisella novicida) response to starvation and oxidative stress. *J. Bacteriol.*, 189:6580–6586, Sep 2007.

[GRG+99] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*, 17(10):994–999, Oct 1999.

[Gus97] D. Gusfield. *Algorithms on strings, trees, and sequences: computer science and computational biology.* Cambridge University Press, New York, NY, USA, 1997.

[GVL96] G.H. Golub and C.F. Van Loan. *Matrix computations*. Johns Hopkins University Press, 1996.

[Han03] S. Hanash. Disease proteomics. *Nature*, 422(6928):226–232, Mar 2003.

[Hay96] M.H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, Inc. New York, NY, USA, 1996.

[HB06] A. Haoudi and H. Bensmail. Bioinformatics and data mining in proteomics. *Expert Rev Proteomics*, 3(3):333–343, Jun 2006.

[HCMB05] J. Hu, K.R. Coombes, J.S. Morris, and K.A. Baggerly. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief Funct Genomic Proteomic*, 3(4):322–331, Feb 2005.

[Hil07] Joseph M. Hilbe. *Negative binomial regression*. New York : Cambridge University Press, 2007.

[HKPM06] M. Hilario, A. Kalousis, C. Pellegrini, and M. Muller. Processing and classification of protein mass spectra. *Mass Spectrometry Reviews*, 25(3), 2006.

[HMA06] P. Hernandez, M. Muller, and R.D. Appel. Automated protein identification by tandem mass spectrometry: Issues and strategies. *Mass Spectrometry Reviews*, 25(2), 2006.

[HMF+05] Y. Hu, J.P. Malone, A.M. Fagan, R.R. Townsend, and D.M. Holtzman. Comparative proteomic analysis of intra- and interindividual variation in human cerebrospinal fluid. *Mol Cell Proteomics*, 4(12):2000–2009, Dec 2005.

[HNR02] C.A. Hastings, S.M. Norton, and S. Roy. New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data. *Rapid Communications in Mass Spectrometry*, 16(5):462–467, 2002.

[HPK07] F. Hillenkamp and J. Peter-Katalinić. *MALDI MS: a practical guide to instrumentation, methods and applications*. Wiley-VCH, 2007.

[HVD08] N.D. Ho and P. Van Dooren. Non-negative matrix factorization with fixed row and column sums. *Linear Algebra and its Applications*, 429(5-6):1020–1025, 2008.

[HWC+03] B.A. Howard, M.Z. Wang, M.J. Campa, C. Corro, M.C. Fitzgerald, and E.F. Patz Jr. Identification and validation of a potential lung cancer serum biomarker detected by matrix-assisted laser desorption/ionization-time of flight spectra analysis. *Proteomics*, 3(9), 2003.

[HWI81] P.J. Huber, J. Wiley, and W. InterScience. *Robust statistics*. Wiley New York, 1981.

[Ind89] Burle Industries. Photomultiplier Handbook. *Inc., Lancaster, Pa*, page 180, 1989.

[JDTP05] R.S. Johnson, M.T. Davis, J.A. Taylor, and S.D. Patterson. Informatics for protein identification by mass spectrometry. *Methods*, 35(3):223–236, 2005.

[Jef05] N. Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21(14):3066–3073, Jul 2005.

[JML+06] J.D. Jaffe, D.R. Mani, K.C. Leptos, G.M. Church, M.A. Gillette, and S.A. Carr. Pepper, a platform for experimental proteomic pattern recognition. *Mol Cell Proteomics*, 5(10):1927–1941, Oct 2006.

[JMP+06] N. Jaitly, M. E. Monroe, V.A Petyuk, T.R.W. Clauss, J.N. Adkins, and R.D. Smith. Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal Chem*, 78(21):7397–7409, Nov 2006.

[JPBF+04] D. Jaitly, R. Page-Belanger, D. Faubert, P. Thibault, and P. Kebarle. MSMS Peak Identification and its Applications. *ISMB/ECCB*, 2004:1–3, 2004.

[JR04] S. Julka and F. Regnier. Quantification in proteomics through stable isotope coding: a review. *J Proteome Res*, 3(3):350–363, 2004.

[JR05] S. Julka and F.E. Regnier. Recent advancements in differential proteomics based on stable isotope coding. *Brief Funct Genomic Proteomic*, 4(2):158–177, Jul 2005.

[JXZ+08] G. Jin, X. Xue, F. Zhang, X. Zhang, Q. Xu, Y. Jin, and X. Liang. Prediction of retention times and peak shape parameters of unknown compounds in traditional Chinese medicine under gradient conditions by ultra performance liquid chromatography. *Analytica Chimica Acta*, 628(1):95–103, 2008.

[KC02] A.N. Krutchinsky and B.T. Chait. On the nature of the chemical noise in MALDI mass spectra. *Journal of the American Society for Mass Spectrometry*, 13(2):129–134, 2002.

[KEH+08] J. Klimek, J.S. Eddes, L. Hohmann, J. Jackson, A. Peterson, S. Letarte, P.R. Gafken, J.E. Katz, P. Mallick, H. Lee, et al. The Standard Protein Mix Database: A Diverse Data Set To Assist in the Production of Improved Peptide and Protein Identification Software Tools. *J. Proteome Res*, 7(01):96–103, 2008.

[KGWR03] J. Kast, M. Gentzel, M. Wilm, and K. Richardson. Noise filtering techniques for electrospray quadrupole time of flight mass spectra. *Journal of the American Society for Mass Spectrometry*, 14(7):766–776, 2003.

[KHO04] J. Karhunen, A. Hyvärinen, and E. Oja. *Independent component analysis*. Wiley-Interscience, 2004.

[KHS+07] Y.V. Karpievitch, E.G. Hill, A.J. Smolka, J.S. Morris, K.R. Coombes, K.A. Baggerly, and J.S. Almeida. PrepMS: TOF MS data graphical preprocessing tool. *Bioinformatics*, 23(2):264, 2007.

[KNKA02] A. Keller, A.I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal Chem*, 74(20):5383–5392, Oct 2002.

[KO05] M. Katajamaa and M. Oresic. Processing methods for differential analysis of LC/MS profile data. *BMC bioinformatics*, 6(1):179, 2005.

[KO07] M. Katajamaa and M. Oresic. Data processing for mass spectrometry-based metabolomics. *J Chromatogr A*, Epub ahead of print, Apr 2007. Epub ahead of print.

[KPRH05] A. Kalousis, J. Prados, E. Rexhepaj, and M. Hilario. Feature extraction from mass spectra for classification of pathological states. *Lecture notes in computer science*, 3721:536, 2005.

[KRR+03] T. Kislinger, K. Rahman, D. Radulovic, B. Cox, J. Rossant, and A. Emili. Prism, a generic large scale proteomic investigation strategy for mammals. *Mol Cell Proteomics*, 2(2):96–106, Feb 2003.

[KSBR04] M. Kempka, J. Sjodahl, A. Bjork, and J. Roeraade. Improved method for peak picking in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 18(11), 2004.

[KSS+07] M. Kirchner, B. Saussen, H. Steen, J.A.J. Steen, and F.A. Hamprecht. amsrpm: Robust point matching for retention time aligment of lc/ms data with r. *Journal of Statistical Software*, 18(4):1–12, 2007.

[KSYW08] B.O. Keller, J. Sui, A.B. Young, and R.M. Whittal. Interferences and contaminants encountered in modern mass spectrometry. *Analytica Chimica Acta*, 627(1):71–81, 2008.

[KTZ+06] M.D. Krebs, R.D. Tingley, J.E. Zeskind, E. Holmboe, J.-M. Kang, and C.E. Davis. Alignment of gas chromatography-mass spectrometry data by landmark selection from complex chemical mixtures. *Chemometrics and Intelligent Laboratory Systems*, 81(1):74–81, March 2006.

[Kut98] Y.A. Kutoyants. *Statistical inference for spatial Poisson processes*. Springer, 1998.

[Lan05] CS Lane. Mass spectrometry-based proteomics in the life sciences. *Cellular and Molecular Life Sciences (CMLS)*, 62(7):848–869, 2005.

[Law87] J.F. Lawless. Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 15(3):209–225, 1987.

[LAYI01] D. Lin, A.J. Alpert, and J.R. Yates III. Multidimensional protein identification technology as an effective tool for proteomics. *Am. Genom. Proteom. Technol*, 1:38–46, 2001.

[LE05] J. Listgarten and A. Emili. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics*, 4(4):419–434, Apr 2005.

[LFP03] L.A. Liotta, M. Ferrari, and E. Petricoin. Clinical proteomics: written in blood. *Nature*, 425(6961):905, 2003.

[LGL+05] X. Li, R. Gentleman, X. Lu, Q. Shi, JD Iglehart, L. Harris, and A. Miron. Seldi-tof mass spectrometry protein data. *Gentleman, R. et al*, pages 91–109, 2005.

[LGR+06] E. Lange, C. Gropl, K. Reinert, O. Kohlbacher, and A. Hildebrandt. High-accuracy peak picking of proteomics data using wavelet techniques. In *Pac. Symp. Biocomput*, volume 11, pages 243–254, 2006.

[LGST+07] E. Lange, C. Gropl, O. Schulz-Trieglaff, A. Leinenbach, C. Huber, and K. Reinert. A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics*, 23(13):i273–i281, Jul 2007.

[Li02] J. Li. Comparison of the capability of peak functions in describing real chromatographic peaks. *Journal of Chromatography A*, 952(1-2):63–70, 2002.

[LKP+03] Q. Liu, B. Krishnapuram, P. Pratapa, X. Liao, A. Hartemink, and L. Carin. Identification of differentially expressed proteins using MALDI-TOF mass spectra. In *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on*, volume 2, 2003.

[LMP00] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multivariee (Troisieme ed.)*. Paris: Dunod, 2000.

[LNR+07] J Listgarten, R.M. Neal, S.T. Roweis, P. Wong, and A. Emili. Difference detection in lc-ms data for protein biomarker discovery. *Bioinformatics*, 23(2):e198–204, 2007.

[LNRE05] J. Listgarten, R.M. Neal, S.T. Roweis, and A. Emili. Multiple alignment of continuous time series. *Advances in Neural Information Processing Systems*, 17:817–824, 2005.

[LOW+05] J. Li, R. Orlandi, C.N. White, J. Rosenzweig, J. Zhao, E. Seregni, D. Morelli, Y. Yu, X.Y. Meng, Z. Zhang, et al. Independent validation of candidate breast cancer serum biomarkers identified by mass spectrometry. *Clinical Chemistry*, 51(12):2229–2235, 2005.

[LSG03] J. Li, H. Steen, and S.P. Gygi. Protein profiling with cleavable isotope-coded affinity tag (cicat) reagents: the yeast salinity stress response. *Mol Cell Proteomics*, 2(11):1198–1204, Nov 2003.

[LSJ+06] K.C. Leptos, D.A. Sarracino, J.D. Jaffe, B. Krastins, and G.M. Church. MapQuant: open-source software for large-scale protein quantification. *Proteomics*, 6(6):1770–1782, 2006.

[LSYr04] H. Liu, RG Sadygov, and JR Yates 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem*, 76(14):4193–201, 2004.

[LTT08] S. Li-Thiao-Té. Semiparametric estimation of the gain parameter with quantization errors. `http://fr.arxiv.org/abs/0906.0346`, 2008.

[LTTS09] S. Li-Thiao-Té and B. Schwikowski. Feature detection with controlled error rates in lc/ms images. *Journal of Computational Biology*, 2009. Submitted.

[LW05] J. Lyons-Weiler. Standards of excellence and open questions in cancer biomarker research: An informatics perspective. *Cancer Informatics*, 1(1):1–7, 2005.

[LYK+05] X.-J. Li, E.C. Yi, C.J. Kemp, H. Zhang, and R. Aebersold. A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol Cell Proteomics*, 4(9):1328–1340, Sep 2005.

[MAB+07] H. Mischak, R. Apweiler, R.E. Banks, M. Conaway, J. Coon, A. Dominiczak, J.H.H. Ehrich, D. Fliser, M. Girolami, H. Hermjakob, D. Hochstrasser, J. Jankowski, B.A. Julian, W. Kolch, Z.A. Massy, C. Neususess, J. Novak, K. Peter, K. Rossing, J. Schanstra, O.J. Semmes, D. Theodorescu, V. Thongboonkerd, E.M. Weissinger, J.E. Van Eyk, and T. Yamamoto. Clinical proteomics: A need to define the field and to begin to set adequate standards. *Proteomics - Clinical Applications*, 1(2):148–156, 2007.

[Man06] M. Mann. Functional and quantitative proteomics using silac. *Nat Rev Mol Cell Biol*, 7(12):952–958, Dec 2006.

[Mar97] R.E. March. An introduction to quadrupole ion trap mass spectrometry. *Journal of Mass Spectrometry*, 32(4):351–369, 1997.

[Mat07] R. Matthiesen. Methods, algorithms and tools in computational proteomics: a practical point of view. *Proteomics*, 7(16), 2007.

[MCA+05] D.I. Malyarenko, W.E. Cooke, B.L. Adam, G. Malik, H. Chen, E.R. Tracy, M.W. Trosset, M. Sasinowski, O.J. Semmes, and D.M. Manos. Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques. *Clinical chemistry*, 51(1):65–74, 2005.

[MCK+05] J.S. Morris, K.R. Coombes, J. Koomen, K.A. Baggerly, and R. Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, 2005.

[McM07] M.C. McMaster. *HPLC: A Practical User's Guide.* J. Wiley & Sons, Inc.: Hoboken, NJ. 2007. xiv + 238pp. ISBN 0-471-75401-3., 2nd ed (university of missouri-st.louis, usa) edition, 2007.

[MCP+03] E. Moskovets, H.S. Chen, A. Pashkova, T. Rejtar, V. Andreev, and B.L. Karger. Closely spaced external standard: a universal method of achieving 5 ppm mass accuracy over the entire MALDI plate in axial matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(19), 2003.

[MCS05] A.J. Matlin, F. Clark, and C.W.J. Smith. Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5):386–398, 2005.

[MH89] R.E. March and R.J. Hughes. *Quadrupole storage mass spectrometry*. Wiley-Interscience, 1989.

[MH02] A.G. Marshall and C.L. Hendrickson. Fourier transform ion cyclotron resonance detection: principles and experimental configurations. *International Journal of Mass Spectrometry*, 215(1-3):59–75, 2002.

[Mil01] X. Milhaud. *Statistique*. Belin, 2001.

[MKG03] A. Menshikov, M. Kleifges, and H. Gemmeke. Fast gain calibration of photomultiplier and electronics. *IEEE Transactions on Nuclear Science*, 50(4 Part 1):1208–1213, 2003.

[MKW+06] AC Miguel, JF Keane, J. Whiteaker, H. Zhang, and A. Paulovich. Compression of LC/MS proteomic data. In *Data Compression Conference, 2006. DCC 2006. Proceedings*, page 1, 2006.

[MMY+06] R.A. Maronna, R.D. Martin, V.J. Yohai, J. Wiley, and W. InterScience. *Robust statistics: theory and methods*. Wiley New York, 2006.

[MN89] P. McCullagh and JA Nelder. Generalized Linear Models, 2nd edition. *Chapman and Hall, London*, 1989.

[MPP+07] D. Mantini, F. Petrucci, D. Pieragostino, P. Del Boccio, M. Di Nicola, C. Di Ilio, G. Federici, P. Sacchetta, S. Comani, and A. Urbani. LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. *BMC Bioinformatics*, 8(1):101, 2007.

[MRS+07] L. Mueller, O. Rinner, A. Schmidt, S. Letarte, B. Bodenmiller, M.Y. Brusniak, O. Vitek, R. Aebersold, and M. Muller. Superhirn – a novel tool for high resolution lc-ms based peptide/protein profiling. *Proteomics*, page in press, 2007. in press.

[MTJ+07] M.E. Monroe, N. Tolic, N. Jaitly, J.L. Shaw, J.N. Adkins, and R.D. Smith. Viper: an advanced software package to support high-throughput lc-ms peptide identification. *Bioinformatics*, epub, Jun 2007.

[MW04] J. Møller and R.P. Waagepetersen. *Statistical inference and simulation for spatial point processes*. Chapman & Hall/CRC, 2004.

[MWS+07] F. Meng, M.C. Wiener, J.R. Sachs, C. Burns, P. Verma, C.P Paweletz, M.T. Mazur, E.G. Deyanova, N.A. Yates, and R.C. Hendrickson. Quantitative analysis of complex peptide mixtures using ftms and differential mass spectrometry. *J Am Soc Mass Spectrom*, 18(2):226–233, Feb 2007.

[NAH+02] S.B. Nielsen, J.U. Andersen, P Hvelplund, T.J.D. Jorgensen, M. Sorensen, and S. Tomita. Triply charged bradykinin and gramicidin radical cations: their formation and the selective enhancement of charge-directed cleavage processes. *International Journal of Mass Spectrometry*, 213(2):225–235, February 2002.

[NF07] K. Noy and D. Fasulo. Improved model-based, platform-independent feature extraction for mass spectrometry. *Bioinformatics*, 23(19):2528, 2007.

[NH88] P.J. Naish and S. Hartwell. Exponentially Modified Gaussian functions—A good model for chromatographic peaks in isocratic HPLC? *Chromatographia*, 26(1):285–296, 1988.

[NKKA03] A.I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *ANALYTICAL CHEMISTRY-WASHINGTON DC-*, 75(17):4646–4658, 2003.

[NMA+05] A.D. Norbeck, M.E. Monroe, J.N. Adkins, K.K. Anderson, D.S. Daly, and R.D. Smith. The utility of accurate mass and lc elution time information in the analysis of complex proteomes. *J Am Soc Mass Spectrom*, 16(8):1239–1249, Aug 2005.

[NW72] J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135(3):370–384, 1972.

[OBK+02] S.-E. Ong, B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, 1(5):376–386, May 2002.

[OM05] S.-E. Ong and M. Mann. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol*, 1(5):252–262, Oct 2005.

[OMAAW+05] W.M. Old, K. Meyer-Arendt, L. Aveline-Wolf, K.G. Pierce, A. Mendoza, J.R. Sevinsky, K.A. Resing, and N.G. Ahn. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics*, 4(10):1487–1502, Oct 2005.

[PAH+02] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, et al. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359(9306):572–577, 2002.

[PBG+07] C. Paulus, S. Bonnet, L. Gerfault, E. Mery, G. Strubel, F. Ricoul, P. Grangeat, and G. CEA-LETI. Chromatographic alignment combined with chemometrics profile reconstruction approaches applied to LC-MS data. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 5983–5986, 2007.

[PEH+04] P.G.A. Pedrioli, J.K. Eng, R. Hubley, M. Vogelzang, E.W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R.H. Angeletti, R. Apweiler, et al. A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology*, 22:1459–1466, 2004.

[PET+03] J. Peng, J.E. Elias, C.C. Thoreen, L.J. Licklider, and S.P. Gygi. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (lc/lc-ms/ms) for large-scale protein analysis: the yeast proteome. *J Proteome Res*, 2(1):43–50, 2003.

[PG01] J. Peng and S. P. Gygi. Proteomics: the move to mixtures. *J Mass Spectrom*, 36(10):1083–1091, Oct 2001.

[PHWA06] P.M. Palagi, P. Hernandez, D. Walther, and R.D. Appel. Proteome informatics i: bioinformatics tools for processing experimental data. *Proteomics*, 6(20):5435–5444, Oct 2006.

[PI06] D.R.A. Prieto and H.J. Issaq. Diagnostic Proteomics. *Proteomics for Biological Discovery*, page 247, 2006.

[PKF+03] K. Petritis, L.J. Kangas, P.L. Ferguson, G.A. Anderson, L. Pasa-Tolic, M.S. Lipton, K.J. Auberry, E.F Strittmatter, Y. Shen, R. Zhao, and R.D. Smith. Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Anal Chem*, 75(5):1039–1048, Mar 2003.

[PM06] J.T. Prince and E.M. Marcotte. Chromatographic alignment of esi-lc-ms proteomics data sets by ordered bijective interpolated warping. *Anal Chem*, 78(17):6140–6152, Sep 2006.

[PMW+06] A. Prakash, P. Mallick, J. Whiteaker, H. Zhang, A. Paulovich, M. Flory, H. Lee, R. Aebersold, and B.. Schwikowski. Signal maps for mass spectrometry-based comparative proteomics. *Mol Cell Proteomics*, 5(3):423–432, Mar 2006.

[PPCC99] D.N. Perkins, D.J.C. Pappin, D.M. Creasy, and J.S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18), 1999.

[PPH06] P.N Pratapa, E.F. Patz, and A.J. Hartemink. Finding diagnostic biomarkers in proteomic spectra. *Pac Symp Biocomput*, pages 279–290, 2006.

[PPW+07] A. Prakash, B. Piening, J. Whiteaker, H. Zhang, S.A. Shaffer, D. Martin, L. Hohmann, K. Cooke, J.M. Olson, S. Hansen, et al. Assessing bias in experiment design for large scale mass spectrometry-based quantitative proteomics. *Molecular & Cellular Proteomics*, 6(10):1741, 2007.

[QAT+03] Y. Qu, B. Adam, M. Thornquist, J.D. Potter, M.L. Thompson, Y. Yasui, J. Davis, P.F. Schellhammer, L. Cazares, M.A. Clements, et al. Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. *Biometrics*, pages 143–151, 2003.

[RCA+04] T. Rejtar, H. Chen, V. Andreev, E. Moskovets, and B.L. Karger. Increased identification of peptides by enhanced data processing of high-resolution MALDI TOF/TOF mass spectra prior to database searching. *Anal. Chem*, 76(20):6017–6028, 2004.

[Rip04] B.D. Ripley. *Spatial statistics*. Wiley-Interscience, 2004.

[RJR+04] D. Radulovic, S. Jelveh, S. Ryu, T.G. Hamilton, E. Foss, Y. Mao, and A. Emili. Informatics Platform for Global Proteomic Profiling and Biomarker Discovery Using Liquid Chromatography-Tandem Mass Spectrometry. *Molecular & Cellular Proteomics*, 3(10):984–997, 2004.

[RMH+07] O. Rinner, L.N Mueller, M. Hubalek, M. Müller, M. Gstaiger, and R. Aebersold. An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nat Biotechnol*, 25(3):345–352, Mar 2007.

[RY06] TW Randolph and Y. Yasui. Multiscale processing of mass spectrometry data. *Biometrics*, 62(2):589–597, 2006.

[SAL+02] R.D Smith, G.A Anderson, M.S Lipton, L. Pasa-Tolic, Y. Shen, T.P. Conrads, T.D. Veenstra, and H.R. Udseth. An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics*, 2(5):513–523, May 2002.

[SBM95] M.W. Senko, S.C. Beu, and F.W. McLafferty. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6(4):229–233, 1995.

[SC78] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on accoustics, speech and signal processing*, Assp-26(1):43–49, February 1978.

[SCYI04] R.G. Sadygov, D. Cociorva, and J.R. Yates III. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nature methods*, 1:195–202, 2004.

[SDD+05] J.C. Silva, R. Denny, C.A. Dorschel, M. Gorenstein, I.J. Kass, G.-Z. Li, T. McKenna, M.J. Nold, K. Richardson, P. Young, and S. Geromanos. Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.*, 77(7):2187–2200, 2005.

[SDM+04] G.A. Satten, S. Datta, H. Moura, A.R. Woolfitt, M.G. Carvalho, G.M. Carlone, B.K. De, A. Pavlopoulos, and J.R. Barr. Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. *Bioinformatics*, 20:17, 2004.

[Ser82] J. Serra. Image analysis and mathematical morphology, vols. 1, 2, 1982.

[Ser88] J. Serra. Image Analysis and Mathematical Morphology. 11: Theoretical Advances, 1988.

[SK83] D. Sankoff and J.B. Kruskal. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* 1983.

[SKBM04] H. Shin, J. Koomen, KA Baggerly, and MK Markey. Towards a noise model of MALDI TOF spectra. *American Association for Cancer Research (AACR) advances in proteomics in cancer research*, 2004.

[SKG97] L.R. Snyder, J.J. Kirkland, and J.L. Glajch. *Practical HPLC Method Development.* Wiley Interscience, 2nd edition edition, 1997.

[SM06] H. Shin and M.K. Markey. A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples. *Journal of Biomedical Informatics*, 39(2):227–248, 2006.

[Smi04] R.M. Smith. *Understanding mass spectra: a basic approach*. Wiley-Interscience, 2004.

[SMKM07] H. Shin, M. Mutlu, J.M. Koomen, and M.K. Markey. Parametric power spectral density analysis of noise from instrumentation in maldi tof mass spectrometry. *Cancer Informatics*, 3:317–328, 2007.

[SOB+06] D.J. States, G.S. Omenn, T.W. Blackwell, D. Fermin, J. Eng, D.W. Speicher, and S.M. Hanash. Challenges in deriving high-confidence protein identifications from data gathered by a hupo plasma proteome collaborative study. *Nat Biotechnol*, 24(3):333–338, Mar 2006.

[SRDM03] R. Simon, M.D. Radmacher, K. Dobbin, and L.M. McShane. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*, 95(1):14–18, Jan 2003.

[SRS03] E.F. Strittmatter, N. Rodriguez, and R.D. Smith. High mass measurement accuracy determination for proteomics using multivariate regression fitting: application to electrospray ionization time-of-flight mass spectrometry. *Anal. Chem*, 75(3):460–468, 2003.

[SS04] A.C. Sauve and T.P. Speed. Normalization, baseline correction and alignment of high-throughput mass spectrometry data. *Proceedings Gensips*, 2004.

[STHG+07] O. Schulz-Trieglaff, R. Hussong, C. Gropl, A. Hildebrandt, and K. Reinert. A fast and accurate algorithm for the quantification of peptides from mass spectrometry data. *Lecture Notes in Computer Science*, 4453:473, 2007.

[STL+06] R. Stolt, R.J.O. Torgrip, J. Lindberg, L. Csenki, J. Kolmert, I. Schuppe-Koistinen, S.P. Jacobsson, K. Boehm, J. Rösgen, HJ Hinz, et al. Second-order peak detection for multicomponent high-resolution LC/MS data. *Analytical Chemistry-Columbus*, 78(4):975–983, 2006.

[STPG+08] O. Schulz-Trieglaff, N. Pfeifer, C. Gröpl, O. Kohlbacher, and K. Reinert. LC-MSsim – a simulation software for liquid chromatography mass spectrometry data. *BMC bioinformatics*, 9(1):423, 2008.

[Str88] L. Stryer. Biochemistry. WH. *Freeman and Co., New York*, 331:348–547, 1988.

[Str08] G. Strubel. *Reconstruction de profils moléculaires : modélisation et inversion d'une chaîne de mesure protéomique*. PhD thesis, Institut Polytechnique de Grenoble, 2008.

[SV04] L. Sleno and D.A. Volmer. Ion activation methods for tandem mass spectrometry. *Journal of Mass Spectrometry*, 39(10):1091–1112, 2004.

[SWB03] G.L. Sacks, C.J. Wolyniak, and J.T. Brenna. Analysis of quantization error in high-precision continuous-flow isotope ratio mass spectrometry. *Journal of Chromatography A*, 1020(2):273–282, 2003.

[SZB94] T. Stephan, J. Zehnpfenning, and A. Benninghoven. Correction of dead time effects in time-of-flight mass spectrometry. *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, 12:405, 1994.

[Tan82] H.H. Tan. A statistical model of the photomultiplier gain process with applications to optical pulse detection. *The Telecommunications and Data Acquisition Progress Report 42-68, January and February 1982*, pages 55–67, 1982.

[TBN08] R. Tautenhahn, C. Bottcher, and S. Neumann. Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics*, 9(1):504, 2008.

[THN+04] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q.T. Le. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics*, 20(17):3034–3044, 2004.

[THNC02] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–6572, May 2002.

[TPS04] K. Tang, J.S. Page, and R.D. Smith. Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 15(10):1416–1423, 2004.

[TvdBA04] G. Tomasi, F. van den Berg, and C. Andersson. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemom.*, 18:231–241, 2004.

[VDY01] S. Veeser, M. J. Dunn, and G. Z. Yang. Multiresolution image registration for two-dimensional gel electrophoresis. *Proteomics*, 1(7):856–870, Jul 2001.

[VJB08] D. Valkenborg, I. Jansen, and T. Burzykowski. A Model-Based Method for the Prediction of the Isotopic Distribution of Peptides. *Journal of the American Society for Mass Spectrometry*, 19(5):703–712, 2008.

[VLTTK+08] M. Vandenbogaert, S. Li-Thiao-Té, H. M. Kaltenbach, R. Zhang, T. Aittokallio, and B. Schwikowski. Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics*, 8:650–672, Feb 2008.

[vNXL+06] A. M. van Nederkassel, C. J. Xu, P. Lancelin, M. Sarraf, D. A. Mackenzie, N. J. Walton, F. Bensaid, M. Lees, G. J. Martin, J. R. Desmurs, D. L. Massart, J. Smeyers-Verbeke, and Y. Vander Heyden. Chemometric treatment of vanillin fingerprint chromatograms. effect of different signal alignments on principal component analysis plots. *J Chromatogr A*, 1120(1-2):291–298, Jul 2006.

[VR02] W.N. Venables and B.D. Ripley. *Modern applied statistics with S*. Springer, 2002.

[VT01] H.L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I*. Wiley-Interscience, 2001.

[VY04] J.D. Venable and J.R. Yates. Impact of ion trap tandem mass spectra variability on the identification of peptides. *Anal Chem*, 76(10):2928–2937, May 2004.

[WAF+03] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13):1636–1643, Sep 2003.

[WAM+05] D.B. Weatherly, J.A. Astwood, T.A. Minning, C. Cavola, R.L. Tarleton, and R. Orlando. A heuristic method for assigning a false-discovery rate for protein identifications from mascot database search results. *Mol Cell Proteomics*, 4(6):762–772, Jun 2005.

[WBF+02] D.B. Wall, S.J. Berger, J.W. Finch, S.A. Cohen, K. Richardson, R. Chapman, D. Drabble, J. Brown, and D. Gostick. Continuous sample deposition from reversed-phase liquid chromatography to tracks on a matrix-assisted laser desorption/ionization precoated target for the analysis of protein digests. *Electrophoresis*, 23(18), 2002.

[WCD+05] B. Williams, S. Cornett, B. Dawant, A. Crecelius, B. Bodenheimer, and R. Caprioli. An algorithm for baseline correction of MALDI mass spectra. In *Proceedings of the 43rd annual Southeast regional conference-Volume 1*, pages 137–142. ACM New York, NY, USA, 2005.

[Wiz79] J.L. Wiza. Microchannel Plate Detectors. *Nucl. Instrum. Methods*, 162:587–601, 1979.

[WKG04] W.E. Wallace, A.J. Kearsley, and C.M. Guttman. An operator-independent approach to mass spectral peak identification and integration. *ANALYTICAL CHEMISTRY-WASHINGTON DC-*, 76(9):2446–2452, 2004.

[WLJR05] W.E. Wolski, M. Lalowski, P. Jungblut, and K. Reinert. Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants. *BMC bioinformatics*, 6(203):1471–2105, 2005.

[WNP03] M. Wagner, D. Naik, and A. Pothen. Protocols for disease classification from mass spectrometry data. *Proteomics*, 3(9), 2003.

[WPP96] W. Windig, J.M. Phalp, and A.W. Payne. A noise and background reduction method for component detection in liquid chromatography/mass spectrometry. *Anal. Chem*, 68(20):3602–3606, 1996.

[WSDY04] M.C. Wiener, J.R. Sachs, E.G. Deyanova, and N.A. Yates. Differential mass spectrometry: a label-free lc-ms method for finding significant differences in complex peptide and protein mixtures. *Anal. Chem.*, 76(20):6085–6096, Oct 2004.

[WTF+07] P. Wang, H. Tang, M.P. Fitzgibbon, M. Mcintosh, M. Coram, H. Zhang, E. Yi, and R. Aebersold. A statistical method for chromatographic alignment of LC-MS data. *Biostatistics*, 8(2):357, 2007.

[WTZ+06] P. Wang, H. Tang, H. Zhang, J. Whiteaker, A.G. Paulovich, and M. Mcintosh. Normalization regarding non-random missing values in high-throughput mass spectrometry data. In *Proc. Pac. Symp. Biocomput*, volume 11, pages 315–326, 2006.

[WW05] B. Walczak and W. Wu. Fuzzy warping of chromatograms. *Chemometrics and Intelligent Laboratory Systems*, 77(1-2):173–180, May 2005. FESTSCHRIFT HONOURING PROFESSOR D.L. MASSART.

[WWZ+06] G. Wang, W.W. Wu, W. Zeng, C.-L. Chou, and R.-F. Shen. Label-free protein quantification using lc-coupled ion trap or ft mass spectrometry: Reproducibility, linearity, and application with complex proteomes. *J Proteome Res*, 5(5):1214–1223, 2006.

[WZL+03] W. Wang, H. Zhou, H. Lin, S. Roy, T.A. Shaler, L.R. Hill, S. Norton, P. Kumar, M. Anderle, and C.H. Becker. Quantification of Proteins and Metabolites by Mass Spectrometry without Isotopic Labeling or Spiked Standards. *Anal. Chem*, 75(18):4818–4826, 2003.

[WZP+06] X. Wang, W. Zhu, K. Pradhan, C. Ji, Y. Ma, O.J. Semmes, J. Glimm, and J. Mitchell. Feature extraction in the analysis of proteomic mass spectra. *Proteomics*, 6(7), 2006.

[YBV08] X. Ye, J. Blonder, and T.D. Veenstra. Targeted proteomics for validation of biomarkers in clinical samples. *Briefings in Functional Genomics and Proteomics*, 2008.

[YC09] A.K. Yocum and A.M. Chinnaiyan. Current affairs in quantitative targeted proteomics: multiple reaction monitoring-mass spectrometry. *Briefings in Functional Genomics and Proteomics*, 2009.

[YDL+02] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.

[YHY09] C. Yang, Z. He, and W. Yu. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC bioinformatics*, 10(1):4, 2009.

[YLS03] K. Yu, Z. Lu, and J. Stander. Quantile regression: applications and current research areas. *The Statistician*, pages 331–350, 2003.

[YPT+03] Y. Yasui, M. Pepe, M.L. Thompson, B.L. Adam, G.L. Wright, Y. Qu, J.D. Potter, M. Winget, M. Thornquist, and Z. Feng. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, 4(3):449–463, 2003.

[YWH+05] W. Yu, B. Wu, T. Huang, X. Li, K. Williams, and H. Zhao. Statistical methods in proteomics. *Springer Handbook of Engineering Statistics*, page 37pages, 2005.

[YWL+06] W. Yu, B. Wu, N. Lin, K. Stone, K. Williams, and H. Zhao. Detecting and aligning peaks in mass spectrometry data with applications to MALDI. *Computational Biology and Chemistry*, 30(1):27–38, 2006.

[ZAA+05] X. Zhang, J.M. Asara, J. Adamec, M. Ouzzani, and A.K. Elmagarmid. Data pre-processing in liquid chromatography-mass spectrometry-based proteomics. *Bioinformatics*, 21(21):4054–4059, Nov 2005.

[ZBR+95] I.A. Zuleta, G.K. Barbula, M.D. Robbins, O.K. Yoon, and R.N. Zare. Micromachined Bradbury-Nielsen Gates. *Anal. Chem*, 67:3952–3957, 1995.

[ZF03] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21:977–1000, June 2003.

[ZM07] R. Zubarev and M. Mann. On the proper use of mass accuracy in proteomics. *Molecular & Cellular Proteomics*, 6(3):377, 2007.

[ZMQS06] J.S.D. Zimmer, M.E. Monroe, W.J. Qian, and R.D. Smith. Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass spectrometry reviews*, 25(3):450, 2006.

[ZNM04] X. Zhang, D.A. Narcisse, and K.K. Murray. On-line single droplet deposition for MALDI mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 15(10):1471–1477, 2004.

[ZWM+03] W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm, and J.S. Kovach. Detection of cancer-specific markers amid massive mass spectral data. *Proceedings of the National Academy of Sciences*, 100(25):14666–14671, 2003.

[ZYZ03] H. Zhu, C.Y. Yu, and H. Zhang. Tree-based disease classification using protein data. *Proteomics*, 3(9), 2003.

Appendix A

Notions of biochemistry

A.1 Molecular species

This section is devoted to some basic notions of chemistry. It aims to provide answers to the following questions:

- Why are there several types of proteins with the same sequence ?
- What is the mass of a protein ?

The first question is related to the replicates of protein signals in LC/MS data. The second is related to identification, i.e. how to relate protein sequence to the measured m/z ratio.

An atom is made of three types of particles: protons, neutrons and electrons. Protons are positively charged, neutrons are neutral particles, and electrons have a negative charge. Protons and neutrons form the nucleus of the atom, while the electrons form a cloud surrounding the nucleus.

An atom which has the same number of protons and electrons is electrically neutral, otherwise it is called an ion. Atoms are categorized depending on their number of protons and neutrons. All atoms with the same number of protons belong to the same *chemical element*. A chemical element exists in several forms, depending on the number of neutrons in the nucleus, and these are called *isotopic forms*.

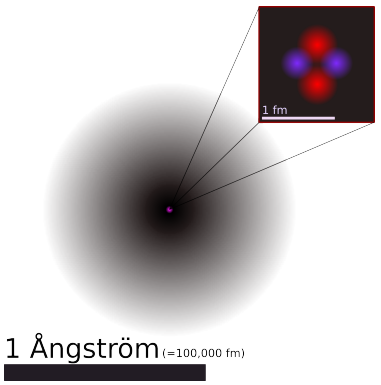

Figure A.1: Atomic structure of the helium atom. The inset shows protons in pink and neutrons in purple.

Atoms are designated with a formula like ^{12}C where C is the chemical element and 12 is the number of protons and neutrons in the nucleus. Atoms with too few or too many neutrons are not stable, but there can exists several stable isotopes of a chemical element. Table A.1 lists common chemical elements in biochemistry, and their isotopes.

Molecular species, rather than molecules, exist in different isotopic forms, depending on the isotopes of their atoms. As isotopes of the same molecule have identical[1] chemical properties, isotopes are distributed in molecules at random. The *mono-isotopic* form of a molecule corresponds to the isotope with all the atoms in their most abundant stable form. It usually corresponds to the lightest isotope form.

[1]Nearly identical. For instance, their molecular weight is different.

Element	Stable isotopes	Relative abundance
Hydrogen	^{1}H	99.985%
	^{2}H	0.015%
Carbon	^{12}C	98.9%
	^{13}C	1.1%
Nitrogen	^{14}N	99.634%
	^{15}N	0.366%
Oxygen	^{16}O	99.76%
	^{17}O	0.039%
	^{18}O	0.201%

Table A.1: Common stable isotopes in protein biochemistry.

The number of additional neutrons is distributed according to a multinomial distribution. It can be approximated with a binomial distribution corresponding to the carbon isotopes as they are the most abundant natural isotopes. Proteins are large molecules with large numbers of carbon atoms, and significant numbers of natural heavy versions. Proteins around 2,000 Da are roughly equally present in the mono-isotopic form and in the isotopic form with an additional neutron.

Molecular weight is a badly defined notion because molecular species exist in forms of different molecular mass. Instead, we use *mono-isotopic mass* which corresponds to the molecular weight of the mono-isotopic form, and average mass which takes into account the stable isotopes. For example, the mono-isotopic mass of Glycine ($C_2H_5NO_2$) is 75 Da, whereas its average mass is 75.03523 Da. A mass spectrometer with 500 ppm accuracy is sufficient to distinguish the two masses.

A.2 From genes to proteins

As described in Chapter 4, protein identification is best performed by comparing fragmentation spectra with the entries of a protein database. In this section, we indicate how to build exhaustive protein databases for MS/MS identification.

The ribosome is the molecular machinery that translates genetic code into proteins. It reads three-letter codons and associates each codon with a specific amino-acid. The association between three-letter codons and amino-acids is fixed in a given organism[2], and it is known for most. Consequently, we can reproduce the translation of the DNA sequence into amino acid sequences on a computer.

To build an exhaustive protein database, it suffices to sequence the genome and apply the translation on a computer. While it may still take months to obtain the complete DNA sequence of complex genomes, simple organisms can be sequenced in a few days.

Remarks

- The amino acid sequence is cut into proteins because some of the three-letter codons indicate the beginning and the termination of a protein sequence.
- The origin of translation is not specified. A given strand of DNA can be read starting from 3 different positions. Additionaly, there are two strands of DNA, and both can be used for coding proteins. However, DNA strands are oriented and can only be read in one way. In total, a DNA sequence needs to be translated six times.

[2]There are variations in the genetic code. For example, some organisms use additional amino-acids like selenocysteine.

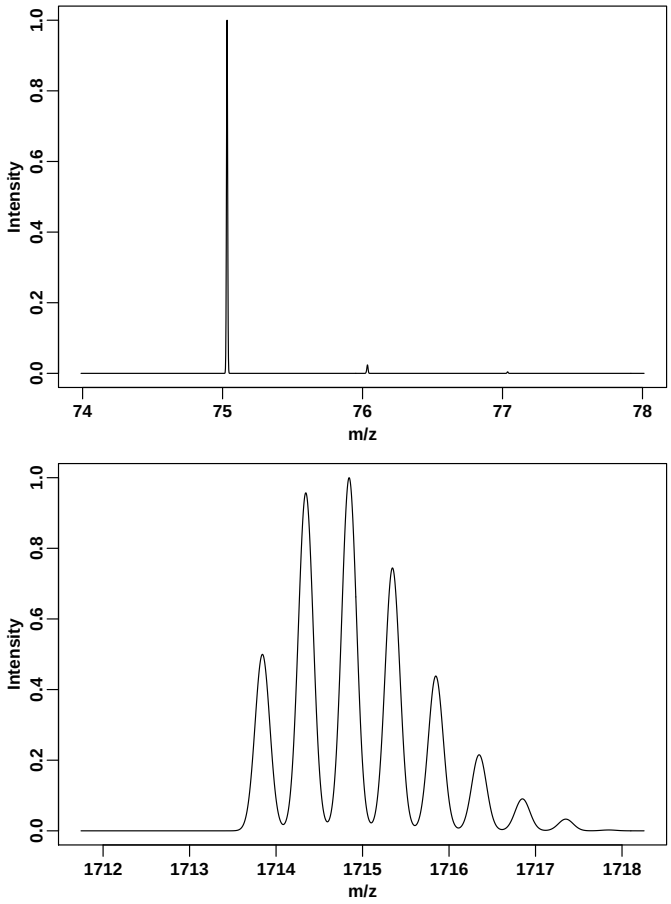

Figure A.2: Isotopic distribution of Glycine (top) and Insulin (bottom) at resolution 20,000. Insulin has two charges, as indicated by the spacing of the peaks.

- There are intermediate steps between the DNA sequence and its translation. Some technologies sequence the intermediate molecules to focus on the genetic information that the cell actually uses.
- There are other biological processes that can increase the complexity of the proteome. For example, splicing corresponds to removing parts of the genetic sequence before translation, and a given sequence can be spliced in different ways (alternative splicing). During maturation, proteins are folded, but also cut. Some parts are removed, e.g. methionines that correspond to the start codon; some proteins are cut into pieces, as is the case for producing hormones.

List of Figures

List of Tables

Traitement du signal et images LC/MS pour la recherche de biomarqueurs.

Dans cette thèse, nous étudions le traitement de données acquises par un spectromètre de masse (MS) couplé à une chromatographie en phase liquide (LC). Cette plateforme de mesure permet d'analyser des mélanges complexes de protéines en séparant les molécules suivant leur masse moléculaire et leur affinité chromatographique.
Les images LC/MS générées souffrent de distortions liées à l'instrument de mesure. Le manuscrit détaille d'abord les caractéristiques des spectromètres usuels en insistant sur les artefacts dans le signal et sur leurs conséquences pour l'identification des protéines, la mesure de leurs concentrations et la recherche de biomarqueurs.

Cette thèse a pour objectif la correction des distortions du signal par des approches mathématiques. Dans un premier temps, nous étudions la non-reproductibilité de la séparation chromatographique et proposons une revue des méthodes de la littérature. Celles-ci sont proches du recalage d'images, avec des contraintes et hypothèses spécifiques aux plateformes LC/MS.
Dans un deuxième temps, nous étudions le bruit de fond dans le signal et comment il affecte la mesure des concentrations de protéines. Pour standardiser les mesures, nous proposons un modèle du bruit de fond basé sur les caractéristiques techniques du spectromètre de masse. En particulier, nous montrons que le gain du détecteur peut être estimé avec une grande précision, et que l'estimateur proposé est optimal.
Pour l'identification des protéines, nous proposons de détecter les signaux dans l'image LC/MS comme des déviations significatives par rapport à un modèle du bruit de fond. Nous considérons un problème de test statistique multiple; nous montrons comment contrôler le taux de faux positifs, le taux de faux négatifs et comment optimiser les paramètres du détecteur.

Signal processing of LC/MS images for biomarker discovery.

This thesis deals with signal processing of images acquired on a mass spectrometer (MS) that is coupled to a liquid chromatography column (LC). In this experimental setup, the molecules in a biological sample are separated in two dimensions, and the instrument records the signal intensity as a function of molecular mass and chromatography position.
LC/MS images are subject to distortions related to the acquisition platform. The manuscript begins with a detailed description of the instruments, artifacts in their measured signal and the consequences for identification of proteins in the sample, quantification and biomarker discovery.

The main topic of the thesis is to correct the signal distortions with mathematical approaches. We first deal with non-reproducibility of the liquid chromatography separation, and propose a review of the available methods in the literature. These methods are similar to multimodal image registration, but have been adapted to the specifics of the LC/MS platform.
Systematic distortions of the measured intensities are then considered, in relation with the background noise in LC/MS images. To standardize the observations, we propose a model for the background noise based on the physical characteristics of the instrument. We show that the gain parameter of the instrument can be estimated with high precision, and that the proposed estimator is optimal.
For identification of the proteins, we propose to detect protein signals in the LC/MS image as significant deviations from the noise model. Feature detection is viewed as a statistical testing problem, which provides a framework for controlling the number of false positive detections, and for measuring the performance of feature detection in terms of test power.