

# 1 Introduction to Mass Spectrometry

“All science is either physics or stamp-collecting.” (Ernest Rutherford)

“The more I learn about experiments, the less I believe in them.” (Cédric Notredame)

GENOMICS AND TRANSCRIPTOMICS have been tremendously successful in the last 40 years. Indispensable prerequisites for this success is our capability to amplify DNA (via polymerase chain reaction) and to read its sequence with low error rates. If we shift to proteomics (the large-scale study of proteins) then it becomes eminent what the analytical problems are: Firstly, there is no way to amplify proteins. Secondly, there is no technique to sequence proteins. Analysis is further complicated by the fact that proteins fold to complex three-dimensional structures, and protein function can only be understood if we take into account these structures. But even if we ignore the latter problem — and we will do so throughout this book — it is understood that our situation is much worse than for Genomics and DNA analysis. This is even more so the case for metabolite and glycan analysis.

Here, mass spectrometry comes into play. This analytical technique has been developed about one century ago, and is routinely applied to biomolecules since the advent of “soft” ionization techniques. Mass spectrometry has many advantages, such as high sensitivity when measuring low-concentration molecules, high speed enabling high-throughput experiments, or high accuracy that allows us to determine the mass of a molecule with outstanding precision. But there is one peculiarity of mass spectrometry that makes its analysis quite different from, say, genome sequencing data: We can only derive a single physical property of the molecules or fragments under consideration, that is, their mass (or, more precisely, their mass-to-charge ratio). Computing the mass of a known molecule is trivial. But how can we get back? How can we make claims about the identity of molecules in our sample, when the only information we have available is the mass of the molecule and, possibly, its fragments?

Computational mass spectrometry is a newly emerged field of research in Bioinformatics with linkage to signal processing, database searching, sequence analysis, statistics, discrete mathematics and graph theory, computational geometry, and others. Computational mass spectrometry addresses the automated analysis of mass spectrometry data, and is of fundamental importance due to the high-throughput nature of mass spectrometry data.

This chapter is meant as a short introduction, to get things started. It is not meant as a reference that you can come back in five to ten years, on a particular subject matter. Instead, it introduces the “bare necessities” from chemistry and mass spectrometry, to provide somebody from bioinformatics or computer science with enough information to dive into the application. Readers with a background in Mass Spectrometry might skip this chapter altogether, and read Chapter 17 instead. Many mass spectrometry and chemistry details have been simplified as much as possible, and you might want to look into the literature to learn what is truly going on — in case this is already known. These simplifications are made to set the focus on the computational side. Also, details can get extremely sophisticated, and easily fill a textbook — and as it turns out, many textbook *have* been filled with these details [49, 53, 102, 146, 214, 215, 233].

## 1.1 Atoms, elements, and molecules

We now give a very short and somewhat oversimplified introduction to atoms and elements. For the moment, we limit ourselves to the pure essentials that are needed to get ourselves going. For more details in particular on isotopes and isotope distributions, we refer to Chapter 9.

Atoms are the building blocks of matter that cannot be decomposed chemically. With the exception of particular environments such as neutron stars, all matter surrounding us is composed of atoms. Atoms in turn are usually composed of three types of massive subatomic particles: *electrons* which have a negative charge, *protons* which have a positive charge, and *neutrons* which have no charge. Protons and neutrons make up the atomic nucleus and are called *nucleons*. Atoms have no charge, and must contain the same number of protons and electrons; if this charge is disrupted, the resulting particle is called an *ion*.

Atoms are classified by their *atomic number*, that is, the number of protons in the atom, that defines which *element* the atom is. All atoms with identical atomic number share the same chemical behavior and cannot be differentiated chemically. The elements most abundant in biomolecules are hydrogen (symbol H) with atomic number 1, carbon (C, atomic number 6), nitrogen (N, 7), oxygen (O, 8), phosphorus (P, 15), and sulfur (S, 16). The “backbone” of all biomolecules is made from carbon. Less abundant elements include fluorine, silicon, copper, zinc, selenium, and tungsten.

Atoms of the same element can differ in their number of neutrons: Such atoms are called isotopes of the element. Different isotopes occur naturally: For example, carbon can have six or seven neutrons, with relative abundance 98.89% and 1.110% in nature, respectively. We will ignore this problem for the moment, and come back to it in Sec. 9.1. For the moment, we assume that all atoms of each element are monoisotopic.

The *mass* of an atom is measured in “unified atomic mass units” with symbol “u”. In biochemistry and molecular biology, the term “Dalton” and the symbol “Da” are used for the same quantity, and we will stick with this notation in the following. In 1961, the International Union of Pure and Applied Chemistry defined 1 Dalton to be 1/12 of the mass of one atom of the carbon-12 isotope.<sup>1</sup> An atom that contains  $n$  protons and neutrons will have a mass of roughly  $n$  Dalton. This is only a rough estimate however, since it does not account for the mass contained in the binding energy of an atom’s nucleus. This explains the *mass defect*, the difference between the atom’s mass and the larger sum of masses of the contained protons, neutrons, and electrons. See Table 1.1 for the monoisotopic masses of the six elements most abundant in living beings. In this book, we will often leave out the unit “Dalton”, in particular in the mathematical context of weighted alphabets. *It should be implicitly understood that all masses in this book are measured in Dalton, unless explicitly stated otherwise.*

A *molecule* consists of a stable system of two or more atoms. Molecules are the smallest particles that retain the chemical properties of the pure chemical substance containing them. The atoms in a molecule are joined by a chemical bond through shared pairs of electrons. The *chemical formula* reflects the exact number of atoms that compose the molecule. A chemical formula may also supply information about the types and spatial arrangement of bonds in the chemical. We use the term *molecular formula* to indicate that we are solely interested in the number of atoms that compose the molecule. Molecules with the same atoms in different arrangements are called *isomers*. For example, the amino acids leucine and isoleucine are

---

<sup>1</sup>Be warned that until 1961, physicists defined 1 amu (atomic mass unit) as 1/16 the mass of one oxygen-16 atom, whereas chemists used the higher *average* mass of oxygen (the atomic weight) as their unit; see Chapter 9.

element	symbol	mass (Da)
hydrogen	H	1.007825
carbon	C	12.0
nitrogen	N	14.003074
oxygen	O	15.994915
phosphor	P	30.973762
sulfur	S	31.972071

Table 1.1: Some important elements with monoisotopic masses in Dalton, rounded to six digits.

isomers. The *nominal mass* (nucleon number) of a molecule is the sum of protons and neutrons of the constituting atoms. The *mass* of a molecule is the sum of masses of the atoms it is composed of.

## 1.2 A tiny primer on biomolecules

This section introduces the “players”: Computational mass spectrometry, to the largest extent, deals with the analysis of biomolecules. For those who are not familiar with this subject matter, we recapitulate some important facts. Everybody else immediately jumps to the next section.

The “manual of life” is written in deoxyribonucleic acids (DNA): it contains the genetic instructions specifying the biological development of all cellular forms of life.<sup>2</sup> A DNA polymer is a chemically linked chain of nucleotides, each of which consists of a sugar, a phosphate and one of four kinds of bases, namely adenine, cytosine, guanine, and thymine. When encoding information, DNA usually appears in the form of a double strand or double helix. The two strands of a DNA double strand usually form a perfect reverse complement of each other. As a macromolecule, a DNA molecule can have a length of several centimeters. The genome of an organism is, roughly speaking, the total information that is encoded in the DNA of its cells. Every cell of an organism carries an identical copy of the genome, with few exceptions such as gametes or mutations.

There exist efficient experimental techniques for analyzing DNA, starting from Polymerase Chain Reaction (PCR) that allows us to replicate DNA at an exponential rate [10], Sanger Sequencing [204] that has been used to sequence the human genome and a few others, Next Generation Sequencing (454 pyrosequencing by Roché, Solexa by Illumina, SOLiD by Applied Biosystems) that can sequence several Gigabases of DNA per day and machine, to third generation methods that are currently being developed. Mass spectrometry never had a dominant role for the analysis of DNA, unlike it has for proteins; with the advent of second and third generation sequencing, this will be even more so. Apart from a few pioneering methods, to which the author of this book has contributed to some extent [23, 66, 147], computational MS does not and will not deal with the analysis of DNA molecules.

Ribonucleic acids (RNA) are biochemically distinguished from DNA by the presence of an additional hydroxyl group, and the use of uracil instead of thymine. A stretch of DNA can be transcribed into RNA, such as messenger RNA (mRNA) encoding proteins. In eukaryotes, certain parts of the RNA molecules are spliced out and the remaining parts are joined, respecting

<sup>2</sup>Biology has few rules without exceptions: RNA viruses encode their genetic instructions in RNA.

the original order. Through this alternative splicing, one DNA sequence can be transcribed into many different mRNA molecules. Other types of RNA exist that do not encode proteins, such as transfer RNA (tRNA) or the famous microRNAs (miRNA). In the lab, there exist several experimental techniques for analyzing RNA, such as microarrays. Regarding RNA analysis and mass spectrometry, the same holds as for DNA and MS.

Finally, the mRNA is translated into a *protein*: Similar to a DNA strand that is a chain of bases, a protein consists of amino acids joined by peptide bonds. An *amino acid* consists of a carboxyl group, an amino group, and the side chain that is specific to each amino acid. Twenty amino acids are encoded by the standard genetic code and are called proteogenic amino acids. Often, amino acids are modified after translation, referred to as Posttranslational Modifications (PTMs). These modifications are not encoded in the DNA or RNA template. The sequence of amino acids constitutes the *primary structure* of the protein. Proteins fold into complex secondary structures (alpha helix, beta sheet) and tertiary structures (spatial relationships in space) that are crucial for their diverse functions. Proteins can also be part of a protein complex, sometimes called quaternary structure. Proteins are essential to the structure and function of all living cells and viruses. Many proteins are enzymes or subunits of enzymes, and catalyze chemical reactions. Other proteins play structural or mechanical roles, are involved in immune response, or the storage and transport of ligands.

Proteins range in size from below 100 amino acids to several thousand amino acids: the muscle protein titin has a single amino acid chain of 27 000 residues. Short sequences of amino acids, as well as parts of digested proteins (see below) are referred to as *peptides*. A variety of PTMs exist in protein biosynthesis, such as the formation of disulfide bridges, or attachment of biochemical functional groups by phosphorylation, acetylation, alkylation and methylation, isoprenylation, glycosylation, and others. The presumably largest PTM is glycosylation, where a (small or large) glycan is attached to one amino acid of the protein, see below.

*Metabolites* are the intermediates and products of metabolism which, in turn, is the entirety of all chemical reactions that happen in living beings to maintain life. Metabolites are rather small, with mass usually below 1000 Dalton. Examples of metabolites are amino acids, monosaccharides, or adenosine-5'-triphosphate (ATP), the energy currency of the cell. Primary metabolites are directly involved in growth, development, and reproduction of a cell or organism; whereas secondary metabolites are not directly involved in those processes. Most of the secondary metabolites in any given higher eukaryote remain unknown. Unlike for proteins, genome sequencing usually does not allow us to deduce the structure of the metabolites. Also unlike proteins or glycans that are made from smaller monomer building blocks, the molecular structure of metabolites is not restricted. This results in a huge variety and complexity of these molecules. We will come back to metabolites in Chapter 13.

There are many sub-classes of metabolites that have particular structural restrictions: For example, *lipids* include fats, waxes, sterols, and fat-soluble vitamins. Lipids may be broadly defined as hydrophobic or amphiphilic small molecules. Other sub-classes of metabolites include nucleotides, amino acids, monosaccharides, steroids, or terpenes. We will not further discuss our utilize the peculiarities of these sub-classes.

*Glycans* are the third major class of biopolymers, and are built from simple sugars (monosaccharides). A large number of monosaccharides exist, but only few are present for an individual species or cell. Glycans can be assembled in a tree-like structure, making their primary structure considerably more complex than that of proteins. Large glycans include starch, cellulose, and chitin. Glycans can be attached to proteins or lipids, but may also be free molecules.

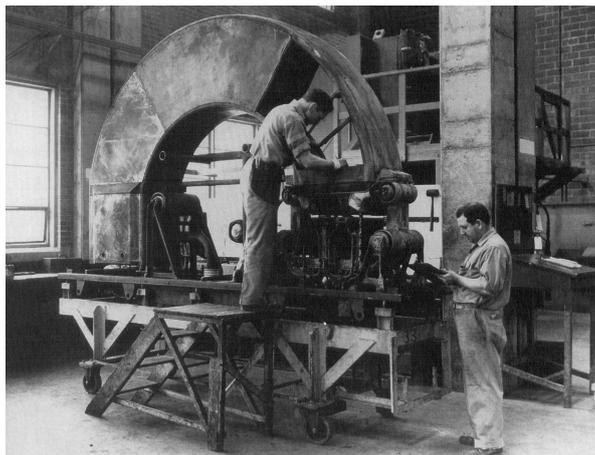


Figure 1.1: The Calutron, used for separating the isotopes of uranium. It was developed by E.O. Lawrence during the Manhattan Project. In 1945, calutrons were used to produce weapons-grade  $^{235}\text{U}$  to build the nuclear bomb [184]. After the war, gaseous diffusion technology and (even later) gas centrifuge technology replaced the calutrons. **[TODO: WEGEN RECHTEN FUER DAS BILD NACHFRAGEN]**

Glycosylation, the attachment of glycans to proteins, is presumably one of the most extensive and complex protein PTM. Glycans are believed to play an important role in cell growth and development, tumor growth and metastasis, or immune recognition and response. Like for metabolites, the structure of glycans cannot be directly inferred from the genome sequence of an organism. We will come back to glycans in Chapter 14.

### 1.3 A short history of Mass Spectrometry

Mass spectrometers constitute a large and diverse class of instruments. Their development began at the end of the 19<sup>th</sup> century with the research on Kanalstrahlen. At that time, experiments by Sir Joseph J. Thomson gave evidence of the existence of stable (non-radioactive) isotopes. In the early 20<sup>th</sup> century, Francis W. Aston (who helped to design Thomson's equipment) and Arthur J. Dempster, along with many others, proved the existence of numerous isotopes, and provided measurements of their abundances and masses. In the following years, mass spectrometry transformed from a somewhat "esoteric" technique into routine instruments used in many labs throughout the world. At the end of World War II, mass spectrometry was used to enrich uranium-235 needed to build the infamous uranium bomb, see Fig. 1.1. The same techniques were later used to provide researchers from physics, chemistry, biology, and medicine with separated isotopes of nearly all the elements of the periodic table. After the war, two important new mass analyzers were invented, namely Time-of-Flight analyzers by William E. Stephens, and quadrupole analyzers by Wolfgang Paul and Hans G. Dehmelt. Also, mass spectrometry was increasingly used to analyze complex organic molecules. Mass spectrometers were coupled to separation techniques such as Gas Chromatography, pioneered by Roland S. Gohlke and Fred W. McLafferty, a combination which is still frequently in use today. Development continued throughout the following years: Tandem mass spectrometry, developed in 1966, coupled two or more mass spectrometers where one machine served as a source for

---

to 1900	early mass spectrometry
1919	observation of isotopes using MS
1946	Time-of-Flight MS (TOF)
1953	quadrupole analyzers
1956	Gas Chromatography MS (GC/MS)
1966	chemical ionization
1966	Tandem MS (MS <sup>2</sup> ), peptide sequencing
1966	metabolomics
1968	Electrospray Ionization (ESI)
1968	Collision Induced Dissociation (CID)
1974	Fourier Transform Ion Cyclotron Resonance (FT-ICR)
1984	Quadrupole/Time-Of-Flight Mass Analyzer (QTOF)
1985	Matrix-Assisted Laser Desorption Ionization (MALDI)
1989	ESI on biomolecules
1992	low level peptide analysis
1993	oligonucleotide ladder sequencing
1993	protein mass mapping
1999	quantitative proteomics and metabolomics with isotope labels
2000	Orbitrap
2004	Electron Transfer Dissociation (ETD)

---

Table 1.2: Short and incomplete list of important developments and inventions in the field of mass spectrometry, tailored towards biomolecules. See the extensive overview at <http://masspec.scripps.edu/mshistory> for much more information.

the next. In the same year, peptide sequencing using mass spectrometry was pioneered by Biemann, Cone, Webster, and Arsenault [20]. Even back then, the interpretation of the mass spectrum was supported by a computer program. Of particular interest for analyzing proteins and other biomolecules is the development of “soft” ionization techniques, namely Electrospray Ionization (ESI) by John B. Fenn, and Matrix-Assisted Laser Desorption Ionization (MALDI) by Michael Karas and Franz Hillenkamp. Fragmenting peptides by Collision Induced Dissociation (CID) was developed in 1968 **[TODO: FORTSETZEN!]** The Orbitrap was invented by Alexander Makarov in 2000 and shows similar performance as an FT-ICR, but without the need of a superconducting magnet. For an overview see Table 1.2.

Six Nobel laureates received their prize for discoveries and inventions in the field of mass spectrometry: Joseph J. (“J.J.”) Thomson (1906, Physics), Francis W. Aston (1922, Chemistry), Wolfgang Paul (1989, Physics), Hans G. Dehmelt (1989, Physics), John B. Fenn (2002, Chemistry), and Koichi Tanaka (2002, Chemistry).

## 1.4 Mass Spectrometry in a Nutshell

To understand what an MS instrument is doing, think of a “parallel scale”. This is a bathroom scale that you can step on, with a twist: Instead of weighting just a single person, we can measure the weights of many people in parallel. Think of a giant scale, where a million people can step on in parallel: Our parallel scale then tells us that 5000 people weighted 64.7 kg, and 12000 people weighted 77.3 kg. We will come back to this example later.

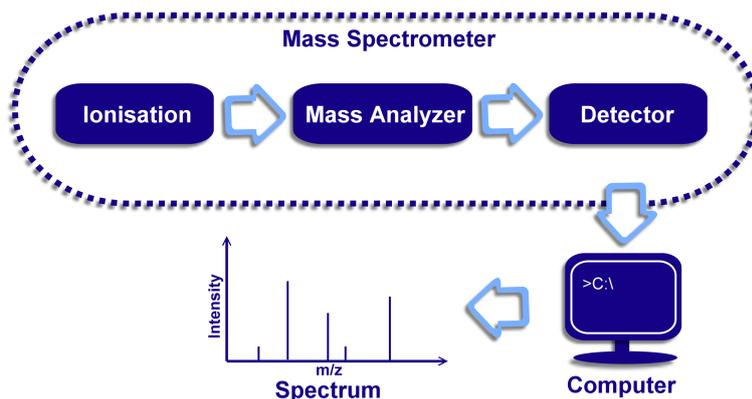


Figure 1.2: Schematic representation of a mass spectrometer.

To analyze molecules in an MS instrument, these molecules first have to be ionized: Only ions will be effected by the electromagnetic field of the mass analyzer. Ionization can happen by attaching one or more protons to the molecule, resulting in a positively charged ion; or, by attaching one or more electrons, resulting in a negatively charged ion. The charge  $z$  of an ion or molecule is a unit-free signed integer  $z \in \mathbb{Z}$ , where unionized molecules have charge  $z = 0$ . Proteins and peptides, which are predominately ionized by ESI or MALDI (see below), are usually positively charged.

It turns out that MS instruments are usually unable to distinguish an ion with mass  $m$  and charge  $z = 1$ , from an ion with mass  $2m$  and charge  $z = 2$  and, more generally, from any ion with mass  $z \cdot m$  and charge  $z \in \mathbb{N}$ . So, we will not be able to record the masses of ions, but only the *mass-to-charge ratio*  $m/z$ . In many applications, we may safely assume that most of the ions have a single charge, and we can treat multiple charged ions separately, without considering them in our general computational setup. For other applications, this is not the case, and we have to deal with different and unknown charges in one measurement. We will come back to this issue in Sec. 7.1. When we talk about recording the masses of ions in our sample, we in fact mean recording the mass-to-charge ratio of the ions, then calculating the masses of the ions under the assumption that, say, all ions are single charged.

Recall that an MS instrument can only analyze ions, whereas uncharged molecules are unaffected by the electromagnetic fields of the instrument and, hence, will not be detected. For convenience, we will often talk about the analyzed molecules, not their charged ion counterparts. Again, it must be understood that we can easily calculate the mass of a molecule if we know the mass of the corresponding ion, as well the “adduct ion” that is responsible for the charge of the ion. Throughout this book, we usually assume that the ion is charged by the addition of a single proton  $H^+$ . In other cases, we can calculate the mass of the molecule analogously, see Sec. 7.3.

In principle, mass spectrometry can be thought of as a three step process: at first, the mixture of molecules that we want to analyze (the *analyte*) has to be ionized in the *ionization source*. Next, analyte ions are separated in the *mass analyzer*. Finally, they hit the *detector* and are recorded. A mass spectrum is then recorded by a computer. These three steps are depicted in Fig. 1.2. Note that this separation into three steps is not true for all types of MS instruments: For example, Orbitrap instruments do not separate mass analyzer and detector. But to understand the principles, it is helpful to think of these steps as being separated; for

the computational analysis, it is usually not important. It should be noted that most MS instruments operate at very low pressure, close to a vacuum, to minimize the random interaction of analyte ions and other particles inside the instrument.

Finally, let us come back to our example of a parallel scale, and talk about some of the limiting factors of an MS analysis. First, assume that our scale tells us that there was a group of people weighting 77.3 kg; but in reality, these people weight 77.4 kg. This corresponds to the *mass accuracy* of the measurement, and is presumably the most important parameters that we have to take into account in our analysis. Next, assume that there are two groups of people, one weighting 88.4 kg, the other 88.6 kg. In this case, our scale might wrongly measure only a single group of people with assumed weight 88.5 kg. This corresponds to the *resolution* of the measurement, and gets important if there are several ions with almost identical mass. Next, certain people might be “easier to detect” for our scale than others, so we cannot trust the numbers of people in each group, and only say that “there is a reasonably large group of people with weight 88.5 kg”. This corresponds to different “ionization preferences” of different molecules.

### [ToDo: PASS OP!]

You will sometimes here people referring to mass spectrometry as “mass spectroscopy”. Using this term is not a good idea, as it might lead to confusion with light spectroscopy — and mass spectrometry has nothing to do with light or radiated energy. Similarly, a mass spectrum has nothing to do with a spectrum of light.

We now describe some ionization sources, mass analyzers, and ion detectors in slightly more detail. This description is again vastly incomplete, and rather meant to introduce some important techniques that one gets in touch with when analyzing MS data. See any MS textbook for more details.

#### 1.4.1 Ionization sources

In the ionization source, analyte molecules are converted into ions. Charge can be created by the addition of removing of a proton, or by adding other adduct ions, see Sec. 7.3. When analyzing biomolecules, the challenge is to create ions without shattering the analyte molecules: In particular proteins are easily fragmented to uninformative pieces using “hard” ionization techniques.

**Electrospray Ionization (ESI)** was developed in 1968 by John B. Fenn. The analyte, dissolved in a large amount of solvent, is pushed through a tiny, highly charged capillary. This liquid pushes itself out of the capillary and forms mist of small droplets. When the solvent evaporates, the analyte molecules are forced closer together; as they have identical charge, they repel each other and eventually break up the droplets. Repeating this process, the droplets get smaller and smaller, until the analyte ions are free of solvent. ESI tends to produce multiple-charged ions, in particular for large analyte molecules such as proteins. See Fig. 1.3 (left), and Fenn *et al.* [73] for more details.

**Matrix-Assisted Laser Desorption/Ionization (MALDI)** was developed in 1985 by Michael Karas and Franz Hillenkamp. The matrix consists of small organic molecules, that absorb energy at the wavelength of the used laser. The method is based on the co-crystallization of the matrix an the analyte components, so that analyte molecules get incorporated into the crystals. The matrix has two functions: it absorbs the light that is fired from the laser, leading to the ionization; and, it protects the molecules of the analyte from

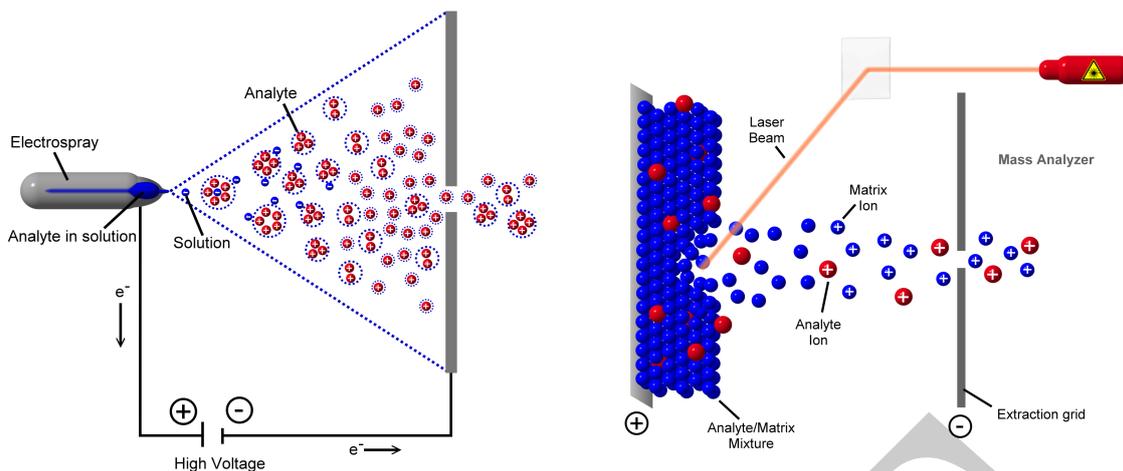


Figure 1.3: Ionization sources: Electrospray ionization (ESI, left) and matrix-assisted laser desorption/ionization (MALDI, right).

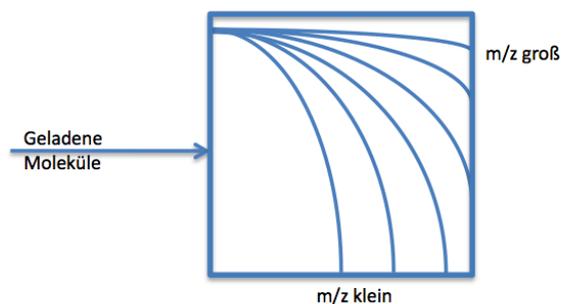


Figure 1.4: Left: Ions in an electromagnetic field. Right: Sector field mass analyzer.

being fragmented by the laser. Very large molecules can be ionized by MALDI without fragmenting them. See Fig. 1.3 (right), and [127] for more details. **[TODO: PASS OP!]**

**Electron Ionization (EI)** was previously called Electron Impact (EI) ionization. It is mainly used in conjunction with Gas Chromatography (see Sec. 1.6.2) for the analysis of small molecules, such as metabolites, see Chapter 13. A beam of energetic electrons is fired at the analyte molecules, inducing ionization and fragmentation. EI is not a “soft” ionization technique, as many of the analyte molecules get fragmented during ionization, often to an extent that no peak is recorded for the mass of the analyte ion. Consequently, EI is practically never used for the analysis of proteins and peptides. But the fragmentation of small molecules is well understood (textbooks have been filled with the details) and the fragmentation spectrum can be used to identify the small molecule by searching in a spectrum library.

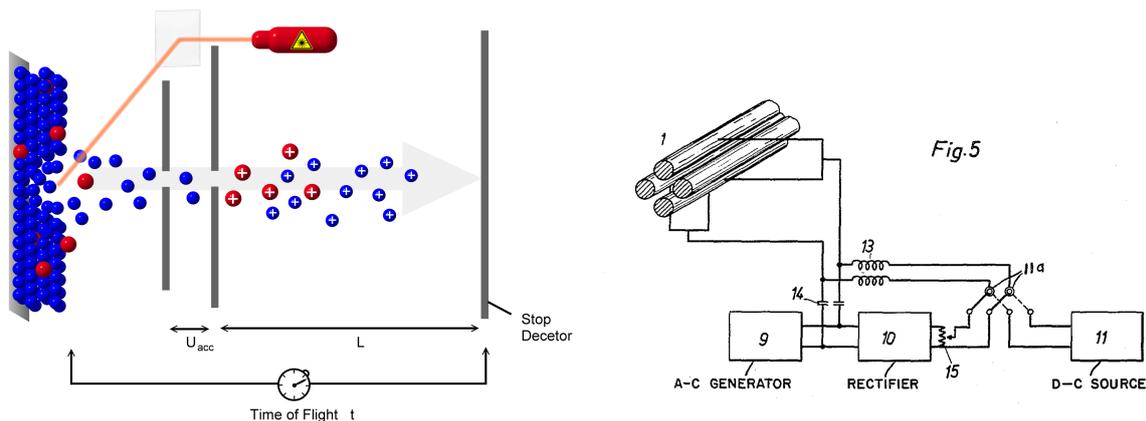


Figure 1.5: Mass analyzers: Quadrupole (left) and Time of Flight (TOF, right). **[TODO: DRAW FIGURES]**

### 1.4.2 Mass analyzers

The second step of the mass spectrometry analysis is presumably the most important part, as this determines the accuracy, sensitivity, resolution, and many other aspects of the machine. All mass analyzers rely on the concept of sending the accelerated ions through an electromagnetic field. In this field, the ions are deflected from their straight line of travel. The higher the charge of the ion, the larger the force of deflection. On the other hand, ions of small mass are easier to deflect than ions of large mass. The conceptually simplest mass analyzer is the sector field mass analyzer, where ions fly in a curved tube; ions can pass the tube if and only if the mass-to-charge ratio fits to the field intensity. By varying the field intensity, we can scan all  $m/z$ -values. See Fig. 1.4.

In the following, we describe four mass analyzers that are, at present, very common for the analysis of biomolecules.

- The **Quadrupole** mass analyzer consists of four circular and parallel rods, that are applied oscillating electric fields. The quadrupole is used to filter analyte ions, based on their mass-to-charge ratio. Only ions of a particular mass-to-charge ratio can pass through the quadrupole on a stable trajectory, see Fig. 1.5. By varying the current applied to the rods, we scan through the range of mass-to-charge ratios. Quadrupole instruments usually achieve rather low mass accuracy, such as 100 parts per million (ppm) or worse. See Miller and Denton [166] for a detailed overview.
- **Time of Flight (TOF)** first accelerates ions in an electric field so that, in principle, all ions have identical kinetic energy. Then, we measure the time ions need to fly through a field-free drift tube, by sampling the current at the detector at discretized time steps. The time-of-flight of an ion depends on its velocity reached during acceleration in the electric field which, in turn, depends on the mass-to-charge ratio of the ions, see Fig. 1.5. Orthogonal acceleration time-of-flight machines can achieve excellent mass accuracy and resolution, and can be regarded as high-accuracy instruments. See Guilhaus [104] for a detailed overview.

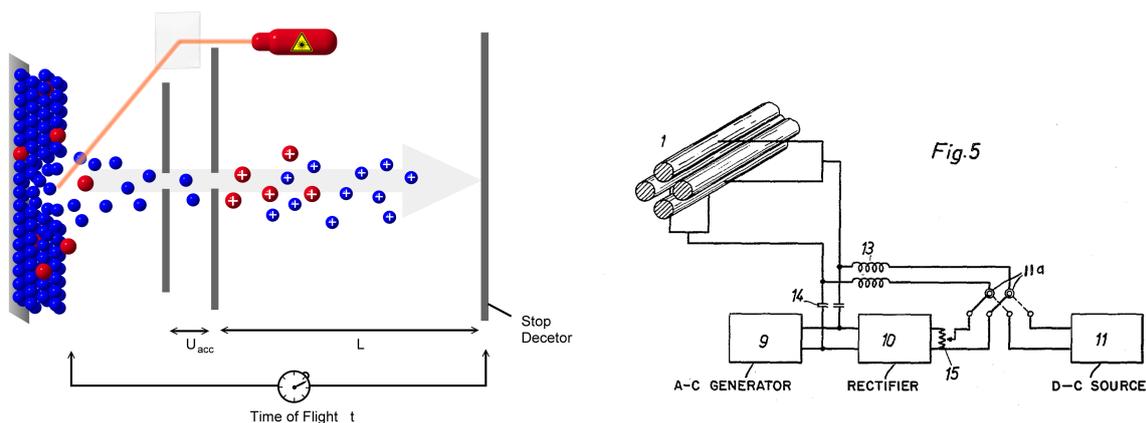


Figure 1.6: Mass analyzers: Fourier Transform Ion Cyclotron Resonance (FT-ICR, bottom left) and Orbitrap (bottom right). **[TODO: DRAW FIGURES]**

- Fourier Transform Ion Cyclotron Resonance (FT-ICR)** These “Penning traps” keep the ions confined in the high magnetic field of a super-conducting magnet. The ions circle with frequencies that are inversely proportional to their  $m/z$  ratio, see Fig. 1.6. This circling induces an alternating current in the metal plates that make up the trap; this time-varying current can be recorded, so FT-ICR does not require a separate ion detector. The current constitutes a frequency spectrum of the ion motion, and is converted into a mass spectrum using the Fourier Transform. FT-ICR instruments have outstanding mass accuracy (sometimes below 1 ppm) and very high resolution.
- The Orbitrap** is an ion trap where moving ions are trapped around an electrode. The electrostatic attraction is compensated by centrifugal force arising from the initial tangential velocity. Potential barriers created by end-electrodes confine the ions axially, see Fig. 1.6. The crux is not so much the analyzer itself, which has been known in Mass Spectrometry for some time, but rather how to get ions into this trap. The LTQ Orbitrap by Thermo Scientific uses several “tricks” to achieve this. The orbitrap mass analyzer is the first fundamentally new mass analyzer introduced commercially in over 20 years. See Hu *et al.* [115], Perry *et al.* [187] for more details.

It must be understood that there is no “perfect” mass analyzer: All come with their particular advantages and disadvantages. For example, quadrupole MS may have a bad mass accuracy; but these instruments are rather cheap, and can be operated at “high” pressure. Also, they can serve as a filter for tandem mass spectrometry, see Section 1.5.

In the MS literature, there is always a race for the best MS instrument; and quite necessarily so, as this drives the development of novel MS methods. But for the computational analysis, we do not have the choice if the instrument that generates our data, is the spearhead in MS instrumentation. Rather, we have to analyze the data at hand; or, estimate what mass accuracy and other parameters are required to get the biological, biochemical, or chemical answers that we are aiming for. The computational analysis of data from an antiquated instrument is obviously a much harder task and, hence, also a bigger intellectual challenge than analyzing data from a top flight instrument. At any given time, 99% of the MS instruments in operation will not be such top flight instruments; so, there is good reason to develop methods for the other

ones, too. In the best case, our computational methods will already work for “low quality” data; but results will continuously get better for data of better quality.

### 1.4.3 Ion detectors

Finally, we have to record the ions that were separated in the mass analyzer. Detectors record either the charge induced or the current produced when an ion passes by or hits a surface. The number of ions that leave the mass analyzer for a particular  $m/z$  value is usually very small, so the signal has to be amplified. Typical ion detectors include electron multiplier, Faraday cups, and microchannel plate detectors. We noted above that for FT-ICR and Orbitrap, the detector is part of the mass analyzer.

It turns out that the actual make of the ion detector is usually not relevant for the computational analysis. Hence, we omit all further details.

## 1.5 Tandem Mass Spectrometry

Tandem mass spectrometry describes numerous techniques where ions with a particular mass-to-charge ratio are selected in a first mass analyzer, are introduced into a fragmentation cell. The most prominent fragmentation technique is *Collision Induced Dissociation* (CID), where molecules are passed through a collision cell containing some noble gas, such as helium or argon. Fragmentation is achieved by collisions with the neutral atoms of the noble gas.

Let us come back to the most interesting part of tandem MS, the collision cell, where our ions collide with the noble gas. Whereas the picture of flying a space ship into an asteroid field is appealing, it is unfortunately wrong: In fact, the fragmentation is rather a chemical process than a physical one. By colliding with the neutral gas, some of the kinetic energy of the molecule is transferred into internal energy. This energy then triggers a fragmentation pathway which, unfortunately, is usually much more complicated than simply cutting some of the bonds in the molecule.

We call the ion that gets fragmented, the *parent ion* or *precursor ion*; the ions in the fragmentation spectrum are called *daughter ions* or *product ions*. When a single charged parent ion is fragmented, the charge of the ion can stay with either of the (usually two) fragments. In this case, the other fragment is not detectable by the MS instrument, and is called a *neutral loss*. It depends on the size of the fragments and, in particular, their chemical structure, which of the fragments is ionized and which is the neutral loss. As we do not fragment a single ion but instead, billions of identical copies, it is still possible that we can detect both fragments of this fragmentation reaction. In case the fragmented ion is multiple charged, the charges are distributed between its fragments. Again, the distribution of charges to the fragment depends on their size and chemical structure.

In passing, we mention that there exist other fragmentation techniques, such as *electron transfer dissociation* (ETD) or *higher-energy collisional dissociation* (HCD). It must be understood that fragmentation spectra from different fragmentation techniques will look vastly different when fragmenting the same molecule.

Finally, there exist other techniques besides tandem MS that allow us to fragment a molecule: Presumably the most prominent is Electron Ionization, where

## 1.6 Sample preparation and separation

Before some molecules can actually be fed to an MS instrument and analyzed there, some steps are taken to make this analysis as simple as possible. In principle, we could directly feed a sample to the instrument, and try to make sense of the data we collect. But this limits the power of our MS analysis, as we have to deal with contamination during the analysis of the data; contaminant signals can superimpose the true signals, dampening or even completely eliminating them. Clearly, it depends on the biological question what we are interested in, what we consider to be “contaminants”: In a proteomics experiment, all metabolites are considered contaminants, whereas the converse is true for a metabolomics experiment. As both the experimental setups and the computational analysis of the data for these two fields are very different, it is practically impossible to analyze them in one go. But as we will see below, separation is a crucial step for a comprehensive analysis, so enriching the molecules we are interested in and getting rid of all others, is always a good idea.

Furthermore, it is sometimes hard or even impossible to directly analyze the biomolecules at hand: A prominent example are proteins, that are “too large” for MS analysis, at least if we are interested in more than their mass alone. So, proteomics analysis requires us to break the proteins into pieces (peptides) before analyzing them by MS. There are at least two important experimental setups for this task, that will be covered in more detail in Chapter 11.

I refrain from describing the experimental details that are needed, say, to extract proteins from a cell. In most cases, these are not important for the computational analysis. In certain situations, though, the exact biochemistry of this extraction can be relevant: An example are “fixed modifications” of amino acids in proteomics experiments (such as the oxidization of methionine, see Sec. 2.6) which are due to the experimental setup, not biochemical processes in the sample. Usually, we can easily modify the computational approach to take into account such peculiarities, and can safely ignore them when developing our computational methods. But there are exceptions to this rule, and peptide database searching in the presence of variable modifications is a prominent example.

### 1.6.1 Tryptic digestion

For some time, there appeared to be a competition in the MS community, regarding the largest intact protein that could be analyzed by mass spectrometry. Apparently, this competition has come to an end; a possible reason being that the mass of an intact protein does not tell you a lot about the protein. Tandem MS of complete proteins is also not possible in most cases, due to various reasons that are beyond the scope of this book. **[TODO: NACHFRAGEN WARUM!]** So, the proteomics community came up with a trick: Instead of analyzing a complete protein, one first cleaves the protein into shorter pieces, namely peptides, then analyzes these peptides by MS.

Proteins can be cleaved into peptides by chemical or enzymatic methods. To understand enzymatic digestion, note that peptide bonds in proteins are metastable, meaning that they will break spontaneously in the presence of water; but this process is extremely slow. Breaking peptide bonds can be leveraged by proteolytic enzymes such as trypsin, V8, or chymotrypsin. At present, the predominant method for protein cleavage is tryptic digestion: The serine protease trypsin cleaves peptide bonds at the carboxy side of a lysine (K) and arginine (R) residue by hydrolysis, adding a water molecule. This cleavage is inhibited if the lysine or arginine residue

is followed by a proline (P). The result of this cleavage are two peptides with sum formulas equal to that of the initial protein, plus  $H_2O$ .

**[ToDo: PASS OP!]**

### 1.6.2 Separation by chromatography

Mixtures of biomolecules are often too complex to be directly fed into an MS instrument: The abundant analyte ions would make it impossible to detect analyte ions that are less frequent in the sample. Also, the huge number of resulting peaks in the mass spectrum, would severely limit our ability to record tandem mass spectra. To this end, these mixtures are usually separated before feeding them to the MS instrument. Chromatography is the predominant separation technique to be coupled with MS.

**[ToDo: PASS OP!]**

Chromatography can be easily coupled with continuous ionization sources such as ESI and EI; for ionization by MALDI, separated parts of the sample first have to be collected before feeding them to the MS instrument.

## 1.7 Exercises

- 1.1 Write a program that simulates protein cleavage by tryptic digestion.

# Bibliography

- [1] A. Aant. I need a title, quick. **[TODO: REPLACE WITH A REAL CITATION]**, 2101.
- [2] G. Alves, A. Y. Ogurtsov and Y.-K. Yu. RAId\_DbS: peptide identification using database searches with realistic statistics. *Biol. Direct.*, 2:25, 2007.
- [3] S. Andreotti, G. W. Klau and K. Reinert. Antilope – a lagrangian relaxation approach to the *de novo* peptide sequencing problem. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2011. To appear, doi:10.1109/TCBB.2011.59.
- [4] R. Apweiler, H. Hermjakob and N. Sharon. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta*, 1473(1): 4–8, 1999.
- [5] G. Audi, A. Wapstra and C. Thibault. The AME2003 atomic mass evaluation (ii): Tables, graphs, and references. *Nucl. Phys. A*, 729:129–336, 2003.
- [6] J.-M. Autebert, J. Berstel and L. Boasson. Context-free languages and pushdown automata. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 111–174. Springer, 1997.
- [7] V. Bafna and N. Edwards. SCOPE: A probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, 17:S13–S21, 2001.
- [8] D. A. Barkauskas and D. M. Rocke. A general-purpose baseline estimation algorithm for spectroscopic data. *Anal. Chim. Acta*, 657(2):191–197, 2010.
- [9] C. Bartels. Fast algorithm for peptide sequencing by mass spectrometry. *Biomed. Environ. Mass Spectrom.*, 19:363–368, 1990.
- [10] J. M. S. Bartlett and D. Stirling. A short history of the polymerase chain reaction. *Methods Mol. Biol.*, 226:3–6, 2003.
- [11] C. Bauer, R. Cramer and J. Schuchhardt. Evaluation of peak-picking algorithms for protein mass spectrometry. *Methods Mol. Biol.*, 696:341–352, 2011.
- [12] M. Beck, I. M. Gessel and T. Komatsu. The polynomial part of a restricted partition function related to the Frobenius problem. *Electron. J. Comb.*, 8(1):N7, 2001.
- [13] D. E. Beihoffer, J. Hendry, A. Nijenhuis and S. Wagon. Faster algorithms for Frobenius numbers. *Electron. J. Comb.*, 12:R27, 2005.
- [14] C. Benecke, T. Grüner, A. Kerber, R. Laue and T. Wieland. MOlecular Structure GENERation with MOLGEN, new features and future developments. *Anal. Chim. Acta*, 314:141–147, 1995.

## Bibliography

- [15] G. Benson. Composition alignment. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2003)*, volume 2812 of *Lect. Notes Comput. Sc.*, pages 447–461. Springer, 2003.
- [16] M. W. Bern and D. Goldberg. EigenMS: De novo analysis of peptide tandem mass spectra by spectral graph partitioning. In *Proc. of Research in Computational Molecular Biology (RECOMB 2005)*, volume 3500 of *Lect. Notes Comput. Sc.*, pages 357–372. Springer, 2005.
- [17] M. W. Bern and D. Goldberg. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *J. Comput. Biol.*, 13(2):364–378, 2006.
- [18] A. Bertsch, A. Leinenbach, A. Pervukhin, M. Lubeck, R. Hartmer, C. Baessmann, Y. A. Elnakady, R. Müller, S. Böcker, C. G. Huber, and O. Kohlbacher. De novo peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. *Electrophoresis*, 30(21):3736–3747, 2009.
- [19] K. Biemann, C. Cone and B. R. Webster. Computer-aided interpretation of high-resolution mass spectra. II. Amino acid sequence of peptides. *J. Am. Chem. Soc.*, 88(11):2597–2598, 1966.
- [20] K. Biemann, C. Cone, B. R. Webster and G. P. Arsenault. Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *J. Am. Chem. Soc.*, 88(23):5598–5606, 1966.
- [21] A. Björklund, T. Husfeldt, P. Kaski and M. Koivisto. Fourier meets Möbius: fast subset convolution. In *Proc. of ACM Symposium on Theory of Computing (STOC 2007)*, pages 67–74. ACM Press New York, 2007.
- [22] N. Blow. Glycobiology: A spoonful of sugar. *Nature*, 457(7229):617–620, 2009.
- [23] S. Böcker. Sequencing from compomers: Using mass spectrometry for DNA de-novo sequencing of 200+ nt. *J. Comput. Biol.*, 11(6):1110–1134, 2004.
- [24] S. Böcker and Zs. Lipták. A fast and simple algorithm for the Money Changing Problem. *Algorithmica*, 48(4):413–432, 2007.
- [25] S. Böcker and V. Mäkinen. Combinatorial approaches for mass spectra recalibration. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 5(1):91–100, 2008.
- [26] S. Böcker and F. Rasche. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24:149–155, 2008. Proc. of *European Conference on Computational Biology (ECCB 2008)*.
- [27] S. Böcker, M. Letzel, Zs. Lipták and A. Pervukhin. Decomposing metabolomic isotope patterns. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2006)*, volume 4175 of *Lect. Notes Comput. Sc.*, pages 12–23. Springer, 2006.
- [28] S. Böcker, B. Kehr and F. Rasche. Determination of glycan structure from tandem mass spectra. In *Proc. of Computing and Combinatorics Conference (COCOON 2009)*, volume 5609 of *Lect. Notes Comput. Sc.*, pages 258–267. Springer, 2009.
- [29] S. Böcker, M. Letzel, Zs. Lipták and A. Pervukhin. SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, 2009.

## Bibliography

- [30] S. Böcker, F. Rasche and T. Steijger. Annotating fragmentation patterns. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2009)*, volume 5724 of *Lect. Notes Comput. Sc.*, pages 13–24. Springer, 2009.
- [31] A. Brauer and J. E. Shockley. On a problem of Frobenius. *J. Reine Angew. Math.*, 211: 215–220, 1962.
- [32] R. Breitling, A. R. Pitt and M. P. Barrett. Precision mapping of the metabolome. *Trends Biotechnol.*, 24(12):543–548, 2006.
- [33] K. Q. Brown. *Geometric transforms for fast geometric algorithms*. Report cmucs-80-101, Dept. Comput. Sci., Carnegie-Mellon Univ., Pittsburgh, USA, 1980.
- [34] S. Cappadona, P. Nanni, M. Benevento, F. Levander, P. Versura, A. Roda, S. Cerutti, and L. Pattini. Improved label-free LC-MS analysis by wavelet-based noise rejection. *J Biomed Biotechnol*, 2010:131505, 2010.
- [35] A. Ceroni, K. Maass, H. Geyer, R. Geyer, A. Dell and S. M. Haslam. GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J. Proteome Res.*, 7 (4):1650–1659, 2008.
- [36] D. C. Chamrad, G. Körting, K. Stühler, H. E. Meyer, J. Klose and M. Blüggel. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*, 4:619–628, 2004.
- [37] S. Chattopadhyay and P. Das. The  $K$ -dense corridor problems. *Pattern Recogn. Lett.*, 11 (7):463–469, 1990.
- [38] E. Check. Proteomics and cancer: Running before we can walk? *Nature*, 429:496–497, 2004.
- [39] T. Chen, M.-Y. Kao, M. Tepel, J. Rush and G. M. Church. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.*, 8(3):325–337, 2001. Preliminary version in *Proc. of Symposium on Discrete Algorithms (SODA 2000)*, Association for Computing Machinery, 2000, 389–398.
- [40] W. L. Chen. Chemoinformatics: past, present, and future. *J. Chem. Inf. Model.*, 46(6): 2230–2255, 2006.
- [41] F. Y. Chin, C. A. Wang and F. L. Wang. Maximum stabbing line in 2D plane. In *Proc. of Conf. on Computing and Combinatorics (COCOON 1999)*, volume 1627 of *Lect. Notes Comput. Sc.*, pages 379–388. Springer, 1999.
- [42] H. H. Chou, H. Takematsu, S. Diaz, J. Iber, E. Nickerson, K. L. Wright, E. A. Muchmore, D. L. Nelson, S. T. Warren, and A. Varki. A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. *Proc. Natl. Acad. Sci. U. S. A.*, 95(20):11751–11756, 1998.
- [43] Y. Chu and T. Liu. On the shortest arborescence of a directed graph. *Sci. Sinica*, 14: 1396–1400, 1965.

## Bibliography

- [44] K. R. Clauser, P. Baker and A. L. Burlingame. Role of accurate mass measurement ( $\pm 10$  ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.*, 71(14):2871–2882, 1999.
- [45] C. A. Cooper, E. Gasteiger and N. H. Packer. GlycoMod – a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics*, 1(2):340–349, 2001.
- [46] C. A. Cooper, H. J. Joshi, M. J. Harrison, M. R. Wilkins and N. H. Packer. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res.*, 31(1):511–513, 2003.
- [47] R. Craig and R. C. Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004.
- [48] V. Dančik, T. A. Addona, K. R. Clauser, J. E. Vath and P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry: A graph-theoretical approach. *J. Comput. Biol.*, 6(3/4):327–342, 1999. Preliminary version in *Proc. of Research in Computational Molecular Biology (RECOMB 1999)*, 135–144.
- [49] C. Dass. *Principles and practice of biological mass spectrometry*. John Wiley and Sons, 2001.
- [50] R. Datta and M. W. Bern. Spectrum fusion: using multiple mass spectra for de novo peptide sequencing. *J. Comput. Biol.*, 16(8):1169–1182, 2009.
- [51] J. L. Davison. On the linear diophantine problem of Frobenius. *J. Number Theory*, 48(3): 353–363, 1994.
- [52] M. de Berg, M. van Kreveld, M. Overmars and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer, second edition, 2000.
- [53] E. de Hoffmann and V. Stroobant. *Mass Spectrometry: Principles and Applications*. Wiley-Interscience, third edition, 2007.
- [54] J. R. de Laeter, J. K. Böhlke, P. D. Bièvre, H. Hidaka, H. S. Peiser, K. J. R. Rosman and P. D. P. Taylor. Atomic weights of the elements. Review 2000 (IUPAC technical report). *Pure Appl. Chem.*, 75(6):683–800, 2003.
- [55] E. W. Deutsch, H. Lam and R. Aebersold. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiological Genomics*, 33:18–25, 2008.
- [56] P. A. DiMaggio and C. A. Floudas. De novo peptide identification via tandem mass spectrometry and integer linear optimization. *Anal. Chem.*, 79(4):1433–1446, 2007.
- [57] B. Domon and R. Aebersold. Mass spectrometry and protein analysis. *Science*, 312:212–217, 2006.
- [58] B. Domon and C. E. Costello. A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate J.*, 5:397–409, 1988.
- [59] R. Dondi, G. Fertin and S. Vialette. Complexity issues in vertex-colored graph pattern matching. *J. Discrete Algorithms*, 2010. In press, doi:10.1016/j.jda.2010.09.002.

## Bibliography

- [60] R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer, 1999.
- [61] S. E. Dreyfus and R. A. Wagner. The Steiner problem in graphs. *Networks*, 1(3):195–207, 1972.
- [62] M. Dyer. Approximate counting by dynamic programming. In *Proc. of Symposium on Theory of Computing (STOC 2003)*, pages 693–699, 2003.
- [63] S. R. Eddy. “antedisciplinary” science. *PLoS Comput. Biol.*, 1(1):e6, 2005.
- [64] P. Edman. Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.*, 4:283–293, 1950.
- [65] J. Edmonds. Optimum branchings. *J. Res. Nat. Bur. Stand.*, 71B:233–240, 1967.
- [66] M. Ehrich, S. Böcker and D. van den Boom. Multiplexed discovery of sequence polymorphisms using base-specific cleavage and MALDI-TOF MS. *Nucleic Acids Res.*, 33(4):e38, 2005.
- [67] D. Einstein, D. Lichtblau, A. Strzebonski and S. Wagon. Frobenius numbers by lattice point enumeration. *INTEGERS*, 7(1):#A15, 2007.
- [68] 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, 2007.
- [69] J. E. Elias, F. D. Gibbons, O. D. King, F. P. Roth and S. P. Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.*, 22(2):214–219, 2004.
- [70] J. K. Eng, A. L. McCormack and J. R. Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectr.*, 5:976–989, 1994.
- [71] M. Ethier, J. A. Saba, M. Spearman, O. Krokhin, M. Butler, W. Ens, K. G. Standing, and H. Perreault. Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 17(24):2713–2720, 2003.
- [72] M. Fellows, G. Fertin, D. Hermelin and S. Vialette. Sharp tractability borderlines for finding connected motifs in vertex-colored graphs. In *Proc. of International Colloquium on Automata, Languages and Programming (ICALP 2007)*, volume 4596 of *Lect. Notes Comput. Sc.*, pages 340–351. Springer, 2007.
- [73] J. Fenn, M. Mann, C. Meng, S. Wong and C. Whitehouse. Electrospray ionisation for mass spectrometry of large biomolecules. *Science*, 246:64–71, 1989.
- [74] D. Fenyö and R. C. Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.*, 75(4):768–774, 2003.
- [75] J. Fernández-de-Cossío, L. J. Gonzalez and V. Besada. A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. *Comput. Appl. Biosci.*, 11(4):427–434, 1995.

## Bibliography

- [76] J. Fernández-de-Cossío, J. Gonzalez, T. Takao, Y. Shimonishi, G. Padron and V. Besada. A software program for the rapid sequence analysis of unknown peptides involving modifications, based on MS/MS data. In *ASMS Conf. on Mass Spectrometry and Allied Topics, Slot 074*, 1997.
- [77] J. Fernández-de-Cossío, L. J. Gonzalez, Y. Satomi, L. Betancourt, Y. Ramos, V. Huerta, A. Amaro, V. Besada, G. Padron, N. Minamino, and T. Takao. Isotopica: a tool for the calculation and viewing of complex isotopic envelopes. *Nucleic Acids Res.*, 32(Web Server issue):W674–W678, 2004.
- [78] A. R. Fernie, R. N. Trethewey, A. J. Krotzky and L. Willmitzer. Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.*, 5(9):763–769, 2004.
- [79] H. I. Field, D. Fenyö and R. C. Beavis. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics*, 2(1):36–47, 2002.
- [80] B. Fischer, V. Roth, F. Roos, J. Grossmann, S. Baginsky, P. Widmayer, W. Gruissem, and J. M. Buhmann. NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal. Chem.*, 77(22):7265–7273, 2005.
- [81] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009. Freely available from <http://algo.inria.fr/flajolet/Publications/book.pdf>.
- [82] A. Frank and P. Pevzner. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.*, 15:964–973, 2005.
- [83] A. M. Frank, M. M. Savitski, M. N. Nielsen, R. A. Zubarev and P. A. Pevzner. De novo peptide sequencing and identification with precision mass spectrometry. *J. Proteome Res.*, 6(1):114–123, 2007.
- [84] A. Fürst, J.-T. Clerc and E. Pretsch. A computer program for the computation of the molecular formula. *Chemom. Intell. Lab. Syst.*, 5:329–334, 1989.
- [85] V. A. Fusaro, D. R. Mani, J. P. Mesirov and S. A. Carr. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat. Biotechnol.*, 27(2):190–198, 2009.
- [86] H. Gabow, Z. Galil, T. Spencer and R. Tarjan. Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica*, 6:109–122, 1986.
- [87] M. R. Garey and D. S. Johnson. *Computers and Intractability (A Guide to Theory of NP-Completeness)*. Freeman, New York, 1979.
- [88] J. Gasteiger, W. Hanebeck and K.-P. Schulz. Prediction of mass spectra from structural information. *J. Chem. Inf. Comput. Sci.*, 32(4):264–271, 1992.
- [89] S. P. Gaucher, J. Morrow and J. A. Leary. STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Anal. Chem.*, 72(11):2331–2336, 2000.
- [90] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant. Open mass spectrometry search algorithm. *J. Proteome Res.*, 3:958–964, 2004.

## Bibliography

- [91] P. Gilmore and R. Gomory. Multi-stage cutting stock problems of two and more dimensions. *Oper. Res.*, 13(1):94–120, 1965.
- [92] D. Goldberg, M. Sutton-Smith, J. Paulson and A. Dell. Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics*, 5(4):865–875, 2005.
- [93] D. Goldberg, M. W. Bern, B. Li and C. B. Lebrilla. Automatic determination of O-glycan structure from fragmentation spectra. *J. Proteome Res.*, 5(6):1429–1434, 2006.
- [94] D. Goldberg, M. W. Bern, S. Parry, M. Sutton-Smith, M. Panico, H. R. Morris and A. Dell. Automated N-glycopeptide identification using a combination of single- and tandem-MS. *J. Proteome Res.*, 6(10):3995–4005, 2007.
- [95] D. Goldberg, M. W. Bern, S. J. North, S. M. Haslam and A. Dell. Glycan family analysis for deducing N-glycan topology from single MS. *Bioinformatics*, 25(3):365–371, 2009.
- [96] A. H. Grange, M. C. Zumwalt and G. W. Sovocool. Determination of ion and neutral loss compositions and deconvolution of product ion mass spectra using an orthogonal acceleration time-of-flight mass spectrometer and an ion correlation program. *Rapid Commun. Mass Spectrom.*, 20(2):89–102, 2006.
- [97] N. A. Gray. Applications of artificial intelligence for organic chemistry: Analysis of C-13 spectra. *Artificial Intelligence*, 22(1):1–21, 1984.
- [98] N. A. B. Gray, R. E. Carhart, A. Lavanchy, D. H. Smith, T. Varkony, B. G. Buchanan, W. C. White, and L. Creary. Computerized mass spectrum prediction and ranking. *Anal. Chem.*, 52(7):1095–1102, 1980.
- [99] N. A. B. Gray, A. Buchs, D. H. Smith and C. Djerassi. Computer assisted structural interpretation of mass spectral data. *Helv. Chim. Acta*, 64(2):458–470, 1981.
- [100] H. Greenberg. Solution to a linear diophantine equation for nonnegative integers. *J. Algorithms*, 9(3):343–353, 1988.
- [101] D. H. Greene and D. E. Knuth. *Mathematics for the Analysis of Algorithms*, volume 1 of *Progress in Computer Science and Applied Logic (PCS)*. Birkhäuser Boston, 1990.
- [102] J. Gross. *Mass Spectrometry: A textbook*. Springer, Berlin, 2004.
- [103] K. Grützmann, S. Böcker and S. Schuster. Combinatorics of aliphatic amino acids. *Naturwissenschaften*, 98(1):79–86, 2011.
- [104] M. Guilhaus. Principles and instrumentation in time-of-flight mass spectrometry. *J. Mass Spectrom.*, 30:1519–1532, 1995.
- [105] S. Guillemot and F. Sikora. Finding and counting vertex-colored subtrees. In *Proc. of Symposium on Mathematical Foundations of Computer Science (MFCS 2010)*, volume 6281 of *Lect. Notes Comput. Sc.*, pages 405–416. Springer, 2010.
- [106] C. Hamm, W. Wilson and D. Harvan. Peptide sequencing program. *Comput. Appl. Biosci.*, 2:115–118, 1986.

## Bibliography

- [107] F. Harary, R. W. Robinson and A. J. Schwenk. Twenty-step algorithm for determining the asymptotic number of trees of various species. *J. Austral. Math. Soc.*, 20(Series A): 483–503, 1975.
- [108] M. Havilio, Y. Haddad and Z. Smilansky. Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.*, 75:435–444, 2003.
- [109] M. Heinonen, A. Rantanen, T. Mielikäinen, J. Kokkonen, J. Kiuru, R. A. Ketola and J. Rousu. FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun. Mass Spectrom.*, 22(19):3043–3052, 2008.
- [110] D. W. Hill, T. M. Kertesz, D. Fontaine, R. Friedman and D. F. Grant. Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra. *Anal. Chem.*, 80(14):5574–5582, 2008.
- [111] W. M. Hines, A. M. Falick, A. L. Burlingame and B. W. Gibson. Pattern-based algorithm for peptide sequencing from tandem high energy collision-induced dissociation mass spectra. *J. Am. Soc. Mass Spectrom.*, 3(4):326 – 336, 1992.
- [112] C. A. R. Hoare. FIND (algorithm 65). *Communications of the ACM*, 4:321–322, 1961.
- [113] D. H. Horn, R. A. Zubarev and F. W. McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectr.*, 11:320–332, 2000.
- [114] C. S. Hsu. Diophantine approach to isotopic abundance calculations. *Anal. Chem.*, 56(8): 1356–1361, 1984.
- [115] Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman and R. G. Cooks. The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.*, 40(4):430–443, 2005.
- [116] R. Hussong and A. Hildebrandt. Signal processing in proteomics. *Methods Mol. Biol.*, 604: 145–161, 2010.
- [117] 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, 2006.
- [118] N. Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21(14):3066–3073, 2005.
- [119] R. S. Johnson and J. A. Taylor. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Methods Mol. Biol.*, 146:41–61, 2000.
- [120] R. S. Johnson and J. A. Taylor. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Mol. Biotechnol.*, 22(3):301–315, 2002.
- [121] P. Jones, R. G. Côté, L. Martens, A. F. Quinn, C. F. Taylor, W. Derache, H. Hermjakob, and R. Apweiler. PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.*, 34(Database-Issue):659–663, 2006.

## Bibliography

- [122] H. J. Joshi, M. J. Harrison, B. L. Schulz, C. A. Cooper, N. H. Packer and N. G. Karlsson. Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics*, 4(6):1650–1664, 2004.
- [123] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble and M. J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*, 4(11): 923–925, 2007.
- [124] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, 34:D354–D357, 2006.
- [125] R. Kannan. Lattice translates of a polytope and the Frobenius problem. *Combinatorica*, 12:161–177, 1991.
- [126] E. A. Kapp, F. Schütz, L. M. Connolly, J. A. Chakel, J. E. Meza, C. A. Miller, D. Fenyo, J. K. Eng, J. N. Adkins, G. S. Omenn, and R. J. Simpson. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. *Proteomics*, 5:3475–3490, 2005.
- [127] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 Daltons. *Anal. Chem.*, 60:2299–2301, 1988.
- [128] 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, 2002.
- [129] A. Keller, J. Eng, N. Zhang, X.-J. Li and R. Aebersold. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.*, 1:2005.0017, 2005.
- [130] E. Kendrick. A mass scale based on  $CH_2 = 14.0000$  for high resolution mass spectrometry of organic compounds. *Anal. Chem.*, 35(13):2146–2154, 1963.
- [131] A. Kerber, R. Laue and D. Moser. Ein Strukturgenerator für molekulare Graphen. *Anal. Chim. Acta*, 235:221 – 228, 1990.
- [132] A. Kerber, R. Laue, M. Meringer and C. Rücker. Molecules in silico: The generation of structural formulae and its applications. *J. Comput. Chem. Japan*, 3(3):85–96, 2004.
- [133] S. Kim, N. Gupta and P. A. Pevzner. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.*, 7(8):3354–3363, 2008.
- [134] S. Kim, N. Bandeira and P. A. Pevzner. Spectral profiles, a novel representation of tandem mass spectra and their applications for de novo peptide sequencing and identification. *Mol. Cell. Proteomics*, 8(6):1391–1400, 2009.
- [135] S. Kim, N. Gupta, N. Bandeira and P. A. Pevzner. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteomics*, 8(1):53–69, 2009.

## Bibliography

- [136] T. Kind and O. Fiehn. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, 7(1):234, 2006.
- [137] T. Kind and O. Fiehn. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8:105, 2007.
- [138] H. Kubinyi. Calculation of isotope distributions in mass spectrometry: A trivial solution for a non-trivial problem. *Anal. Chim. Acta*, 247:107–119, 1991.
- [139] K.-S. Kwok, R. Venkataraghavan and F. W. McLafferty. Computer-aided interpretation of mass spectra. III. Self-training interpretive and retrieval system. *J. Am. Chem. Soc.*, 95(13):4185–4194, 1973.
- [140] V. Lacroix, C. G. Fernandes, and M.-F. Sagot. Motif search in graphs: Application to metabolic networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 3(4):360–368, 2006.
- [141] A. J. Lapadula, P. J. Hatcher, A. J. Hanneman, D. J. Ashline, H. Zhang and V. N. Reinhold. Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MS<sup>n</sup> data. *Anal. Chem.*, 77(19):6271–6279, 2005.
- [142] R. L. Last, A. D. Jones and Y. Shachar-Hill. Towards the plant metabolome and beyond. *Nat. Rev. Mol. Cell Biol.*, 8:167–174, 2007.
- [143] A. Lavanchy, T. Varkony, D. H. Smith, N. A. B. Gray, W. C. White, R. E. Carhart, B. G. Buchanan, and C. Djerassi. Rule-based mass spectrum prediction and ranking: Applications to structure elucidation of novel marine sterols. *Org. Mass Spectrom.*, 15(7):355–366, 1980.
- [144] J. Lederberg. Topological mapping of organic molecules. *Proc. Natl. Acad. Sci. U. S. A.*, 53(1):134–139, 1965.
- [145] J. Lederberg. How DENDRAL was conceived and born. In *ACM Conference on the History of Medical Informatics, History of Medical Informatics archive*, pages 5–19, 1987. Available from <http://doi.acm.org/10.1145/41526.41528>.
- [146] T. A. Lee. *A Beginner's Guide to Mass Spectral Interpretation*. Wiley, 1998.
- [147] M. Lefmann, C. Honisch, S. Boecker, N. Storm, F. von Wintzingerode, C. Schloetelburg, A. Moter, D. van den Boom, and U. B. Goebel. A novel mass spectrometry based tool for genotypic identification of mycobacteria. *J. Clin. Microbiol.*, 42(1):339–346, 2004.
- [148] G. Li and F. Ruskey. The advantages of forward thinking in generating rooted and free trees. In *Proc. of ACM-SIAM Symposium on Discrete Algorithms (SODA 1999)*, pages 939–940, Philadelphia, PA, USA, 1999. Society for Industrial and Applied Mathematics.
- [149] G. Liu, J. Zhang, B. Larsen, C. Stark, A. Breitkreutz, Z.-Y. Lin, B.-J. Breitkreutz, Y. Ding, K. Colwill, A. Pasculescu, T. Pawson, J. L. Wrana, A. I. Nesvizhskii, B. Raught, M. Tyers, and A.-C. Gingras. ProHits: integrated software for mass spectrometry-based interaction proteomics. *Nat. Biotechnol.*, 28(10):1015–1017, 2010.

## Bibliography

- [150] K. K. Lohmann and C.-W. von der Lieth. GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res.*, 32(Web Server issue):W261–W266, 2004.
- [151] B. Lu and T. Chen. A suffix tree approach to the interpretation of tandem mass spectra: Applications to peptides of non-specific digestion and post-translational modifications. *Bioinformatics*, 19(Suppl 2):ii113–ii121, 2003. Proc. of *European Conference on Computational Biology (ECCB 2003)*.
- [152] A. Luedemann, K. Strassburg, A. Erban and J. Kopka. TagFinder for the quantitative analysis of gas chromatography–mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics*, 24(5):732–737, 2008.
- [153] G. S. Lueker. Two NP-complete problems in nonnegative integer programming. Technical Report TR-178, Department of Electrical Engineering, Princeton University, 1975.
- [154] Y.-R. Luo. *Handbook of Bond Dissociation Energies in Organic Compounds*. CRC Press, Boca Raton, 2003.
- [155] B. Ma and G. Lajoie. Improving the de novo sequencing accuracy by combining two independent scoring functions in peaks software. Poster at the ASMS Conference on Mass Spectrometry and Allied Topics, 2005.
- [156] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby and G. Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 17(20):2337–2342, 2003.
- [157] B. Ma, K. Zhang and C. Liang. An effective algorithm for peptide de novo sequencing from MS/MS spectra. *J. Comput. Syst. Sci.*, 70:418–430, 2005.
- [158] K. Maass, R. Ranzinger, H. Geyer, C.-W. von der Lieth and R. Geyer. “Glyco-peakfinder” – de novo composition analysis of glycoconjugates. *Proteomics*, 7(24):4435–4444, 2007.
- [159] P. Mallick, M. Schirle, S. S. Chen, M. R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster, and R. Aebersold. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.*, 25(1):125–131, 2007.
- [160] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.*, 66(24):4390–4399, 1994.
- [161] S. Martello and P. Toth. An exact algorithm for large unbounded knapsack problems. *Oper. Res. Lett.*, 9(1):15–20, 1990.
- [162] S. Martello and P. Toth. *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley & Sons, Chichester, 1990.
- [163] R. Matthiesen, J. Bunkenborg, A. Stensballe, O. N. Jensen, K. G. Welinder and G. Bauw. Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics*, 4(9):2583–2593, 2004.
- [164] R. Matthiesen, M. B. Trelle, P. Hojrup, J. Bunkenborg and O. N. Jensen. VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. *J. Proteome Res.*, 4(6):2338–2347, 2005.

## Bibliography

- [165] L. McHugh and J. W. Arthur. Computational methods for protein identification from mass spectrometry data. *PLoS Comput. Biol.*, 4(2):e12, 2008.
- [166] P. E. Miller and M. B. Denton. The quadrupole mass filter: Basic operating concepts. *J. Chem. Educ.*, 63:617–622, 1986.
- [167] L. Mo, D. Dutta, Y. Wan and T. Chen. MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Anal. Chem.*, 79(13):4870–4878, 2007.
- [168] E. Mostacci, C. Truntzer, H. Cardot and P. Ducoroy. Multivariate denoising methods combining wavelets and principal component analysis for mass spectrometry data. *Proteomics*, 10(14):2564–2572, 2010.
- [169] I. K. Mun and F. W. McLafferty. Computer methods of molecular structure elucidation from unknown mass spectra. In *Supercomputers in Chemistry*, ACS Symposium Series, chapter 9, pages 117–124. American Chemical Society, 1981.
- [170] S. Na, J. Jeong, H. Park, K.-J. Lee and E. Paek. Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach. *Mol. Cell. Proteomics*, 7(12): 2452–2463, 2008.
- [171] S. Neumann and S. Böcker. Computational mass spectrometry for metabolomics – a review. *Anal. Bioanal. Chem.*, 398(7):2779–2788, 2010.
- [172] N. Nguyen, H. Huang, S. Oraintara and A. Vo. Mass spectrometry data processing using zero-crossing lines in multi-scale of Gaussian derivative wavelet. *Bioinformatics*, 26(18): i659–i665, 2010.
- [173] R. Niedermeier. *Invitation to Fixed-Parameter Algorithms*. Oxford University Press, 2006.
- [174] J. A. November. *Digitizing life: the introduction of computers to biology and medicine*. PhD thesis, Princeton University, Princeton, USA, 2006.
- [175] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, and H. C. Köfeler. On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. results of an austrian multicenter study. *J. Mass Spectrom.*, 44(4):485–493, 2009.
- [176] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, and H. C. Köfeler. On the inter-instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2. optimization and characterization of the search algorithm. *J. Mass Spectrom.*, 44(4):494–502, 2009.
- [177] S. Orchard, L. Montechi-Palazzi, E. W. Deutsch, P.-A. Binz, A. R. Jones, N. Paton, A. Pizarro, D. M. Creasy, J. Wojcik, and H. Hermjakob. Five years of progress in the standardization of proteomics data: 4th annual spring workshop of the HUPO-proteomics standards initiative. *Proteomics*, 7:3436–3440, 2007.
- [178] R. Otter. The number of trees. *The Annals of Mathematics*, 49(3):583–599, 1948.

## Bibliography

- [179] K. G. Owens. Application of correlation analysis techniques to mass spectral data. *Appl. Spectrosc. Rev.*, 27(1):1–49, 1992.
- [180] N. H. Packer, C.-W. von der Lieth, K. F. Aoki-Kinoshita, C. B. Lebrilla, J. C. Paulson, R. Raman, P. Rudd, R. Sasisekharan, N. Taniguchi, and W. S. York. Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11-13, 2006). *Proteomics*, 8(1):8–20, 2008.
- [181] G. Palmisano, D. Antonacci and M. R. Larsen. Glycoproteomic profile in wine: a ‘sweet’ molecular renaissance. *J. Proteome Res.*, 9(12):6148–6159, 2010.
- [182] D. J. Pappin, P. Hojrup and A. Bleasby. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.*, 3(6):327–332, 1993.
- [183] C. Y. Park, A. A. Klammer, L. Käll, M. J. MacCoss and W. S. Noble. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.*, 7(7):3022–3027, 2008.
- [184] W. E. Parkins. The uranium bomb, the calutron, and the space-charge problem. *Physics Today*, 58(5):45–51, 2005.
- [185] V. Pellegrin. Molecular formulas of organic compounds: the nitrogen rule and degree of unsaturation. *J. Chem. Educ.*, 60(8):626–633, 1983.
- [186] D. N. Perkins, D. J. Pappin, D. M. Creasy and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.
- [187] R. H. Perry, R. G. Cooks and R. J. Noll. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom. Rev.*, 27(6):661–699, 2008.
- [188] G. Pólya. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta Mathematica*, 68(1):145–254, 1937.
- [189] S. C. Pomerantz, J. A. Kowalak and J. A. McCloskey. Determination of oligonucleotide composition from mass spectrometrically measured molecular weight. *J. Am. Soc. Mass Spectrom.*, 4:204–209, 1993.
- [190] R. Raman, S. Raguram, G. Venkataraman, J. C. Paulson and R. Sasisekharan. Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nat. Methods*, 2(11):817–824, 2005.
- [191] J. L. Ramírez-Alfonsín. *The Diophantine Frobenius Problem*. Oxford University Press, 2005.
- [192] J. L. Ramírez-Alfonsín. Complexity of the Frobenius problem. *Combinatorica*, 16(1):143–147, 1996.
- [193] F. Rasche, A. Svatoš, R. K. Maddula, C. Böttcher and S. Böcker. Computing fragmentation trees from tandem mass spectrometry data. *Anal. Chem.*, 83:1243–1251, 2011.
- [194] I. Rauf, F. Rasche and S. Böcker. Computing maximum colorful subtrees in practice. Manuscript. **[TODO: REMOVE OR UPDATE]**, 2011.

## Bibliography

- [195] A. L. Rockwood and P. Haimi. Efficient calculation of accurate masses of isotopic peaks. *J. Am. Soc. Mass Spectrom.*, 17(3):415–419, 2006.
- [196] A. L. Rockwood, M. M. Kushnir and G. J. Nelson. Dissociation of individual isotopic peaks: Predicting isotopic distributions of product ions in MS<sup>n</sup>. *J. Am. Soc. Mass Spectr.*, 14:311–322, 2003.
- [197] A. L. Rockwood, J. R. Van Orman and D. V. Dearden. Isotopic compositions and accurate masses of single isotopic peaks. *J. Am. Soc. Mass Spectr.*, 15:12–21, 2004.
- [198] P. Roepstorff and J. Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.*, 11(11):601, 1984.
- [199] S. Rogers, R. A. Scheltema, M. Girolami and R. Breitling. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, 25(4):512–518, 2009.
- [200] R. G. Sadygov and J. R. Yates III. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.*, 75(15):3792–3798, 2003.
- [201] 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. *Nat. Methods*, 1(3):195–202, 2004.
- [202] T. Sakurai, T. Matsuo, H. Matsuda and I. Katakuse. PAAS 3: A computer program to determine probable sequence of peptides from mass spectrometric data. *Biomed. Mass Spectrom.*, 11(8):396–399, 1984.
- [203] A. Salomaa. Counting (scattered) subwords. *B. Euro. Assoc. Theo. Comp. Sci.*, 81:165–179, 2003.
- [204] F. Sanger, S. Nicklen and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5463–5467, 1977.
- [205] M. M. Savitski, M. L. Nielsen, F. Kjeldsen and R. A. Zubarev. Proteomics-grade de novo sequencing approach. *J. Proteome Res.*, 4:2348–2354, 2005.
- [206] K. Scheubert, F. Hufsky, F. Rasche and S. Böcker. Computing fragmentation trees from metabolite multiple mass spectrometry data. In *Proc. of Research in Computational Molecular Biology (RECOMB 2011)*, 2011. To be presented.
- [207] J. Seidler, N. Zinn, M. E. Boehm and W. D. Lehmann. De novo sequencing of peptides by MS/MS. *Proteomics*, 10(4):634–649, 2010.
- [208] J. Senior. Partitions and their representative graphs. *Am. J. Math.*, 73(3):663–689, 1951.
- [209] B. Shan, B. Ma, K. Zhang and G. Lajoie. Complexities and algorithms for glycan sequencing using tandem mass spectrometry. *J. Bioinformatics and Computational Biology*, 6(1):77–91, 2008.

## Bibliography

- [210] Q. Sheng, Y. Mechref, Y. Li, M. V. Novotny and H. Tang. A computational approach to characterizing bond linkages of glycan isomers using matrix-assisted laser desorption/ionization tandem time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, 22(22):3561–3569, 2008.
- [211] I. V. Shilov, S. L. Seymour, A. A. Patel, A. Loboda, W. H. Tang, S. P. Keating, C. L. Hunter, L. M. Nuwaysir, and D. A. Schaeffer. The paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics*, 6(9):1638–1655, 2007.
- [212] H. Shin, M. P. Sampat, J. M. Koomen and M. K. Markey. Wavelet-based adaptive denoising and baseline correction for MALDI TOF MS. *OMICS*, 14(3):283–295, 2010.
- [213] F. Sikora. An (almost complete) state of the art around the graph motif problem. Technical report, Université Paris-Est, France, 2010. Available from <http://www-igm.univ-mlv.fr/~fsikora/pub/GraphMotif-Resume.pdf>.
- [214] R. M. Silverstein, F. X. Webster and D. Kiemle. *Spectrometric Identification of Organic Compounds*. Wiley, 7<sup>th</sup> edition, 2005.
- [215] G. Siuzdak. *The Expanding Role of Mass Spectrometry in Biotechnology*. MCC Press, second edition, 2006.
- [216] D. H. Smith, N. A. Gray, J. G. Nourse and C. W. Crandell. The DENDRAL project: recent advances in computer-assisted structure elucidation. *Anal. Chim. Acta*, 133(4):471 – 497, 1981.
- [217] R. K. Snider. Efficient calculation of exact mass isotopic distributions. *J. Am. Soc. Mass Spectrom.*, 18(8):1511–1515, 2007.
- [218] H. M. Sobell. Actinomycin and DNA transcription. *Proc. Natl. Acad. Sci. U. S. A.*, 82(16): 5328–5331, 1985.
- [219] H. Steen and M. Mann. The ABC's (and XYZ's) of peptide sequencing. *Nature Rev.*, 5: 699–711, 2004.
- [220] M. T. Sykes and J. R. Williamson. Envelope: interactive software for modeling and fitting complex isotope distributions. *BMC Bioinformatics*, 9:446, 2008.
- [221] J. J. Sylvester and W. J. Curran Sharp. Problem 7382. *Educational Times*, 37:26, 1884.
- [222] D. L. Tabb, M. J. MacCoss, C. C. Wu, S. D. Anderson and J. R. Yates. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal. Chem.*, 75(10):2470–2477, 2003.
- [223] H. Tang, Y. Mechref and M. V. Novotny. Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics*, 21 Suppl 1:i431–i439, 2005. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2005).
- [224] S. Tanner, H. Shu, A. Frank, L.-C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna. Inspect: Identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.*, 77:4626–4639, 2005.

## Bibliography

- [225] J. A. Taylor and R. S. Johnson. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.*, 73(11):2594–2604, 2001.
- [226] J. A. Taylor and R. S. Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 11:1067–1075, 1997.
- [227] J. van Lint and R. Wilson. *A Course in Combinatorics*. Cambridge University Press, 2001.
- [228] A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, G. W. Hart, and M. E. Etzler, editors. *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, second edition, 2009. Freely available from <http://www.ncbi.nlm.nih.gov/books/NBK1908/>.
- [229] R. Venkataraghavan, F. W. McLafferty and G. E. van Lear. Computer-aided interpretation of mass spectra. *Org. Mass Spectrom.*, 2(1):1–15, 1969.
- [230] C.-W. von der Lieth, A. Bohne-Lang, K. K. Lohmann and M. Frank. Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief. Bioinform.*, 5(2):164–178, 2004.
- [231] S. A. Waksman and H. B. Woodruff. Bacteriostatic and bacteriocidal substances produced by soil actinomycetes. *Proc. Soc. Exper. Biol.*, 45:609–614, 1940.
- [232] M. S. Waterman and M. Vingron. Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl. Acad. Sci. U. S. A.*, 91(11):4625–4628, 1994.
- [233] J. T. Watson and O. D. Sparkman. *Introduction to Mass Spectrometry: Instrumentation, Applications, and Strategies for Data Interpretation*. Wiley, 2007.
- [234] M. E. Wieser. Atomic weights of the elements 2005 (IUPAC technical report). *Pure Appl. Chem.*, 78(11):2051–2066, 2006.
- [235] H. Wilf. *generatingfunctionology*. Academic Press, second edition, 1994. Freely available from <http://www.math.upenn.edu/~wilf/DownldGF.html>.
- [236] S. Wolf, S. Schmidt, M. Müller-Hannemann and S. Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11:148, 2010.
- [237] 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, 2005.
- [238] J. W. Wong, G. Cagney and H. M. Cartwright. SpecAlign—processing and alignment of mass spectra datasets. *Bioinformatics*, 21(9):2088–2090, 2005.
- [239] L.-C. Wu, H.-H. Chen, J.-T. Horng, C. Lin, N. E. Huang, Y.-C. Cheng and K.-F. Cheng. A novel preprocessing method using Hilbert Huang transform for MALDI-TOF and SELDI-TOF mass spectrometry data. *PLoS One*, 5(8):e12493, 2010.

## Bibliography

- [240] Y. Wu, Y. Mechref, I. Klouckova, M. V. Novotny and H. Tang. A computational approach for the identification of site-specific protein glycosylations through ion-trap mass spectrometry. In *Proc. of RECOMB 2006 satellite workshop on Systems biology and computational proteomics*, volume 4532 of *Lect. Notes Comput. Sc.*, pages 96–107. Springer, 2007.
- [241] C. Xu and B. Ma. Complexity and scoring function of MS/MS peptide de novo sequencing. In *Proc. of Computational Systems Bioinformatics Conference (CSB 2006)*, volume 4 of *Series on Advances in Bioinformatics and Computational Biology*, pages 361–369. Imperial College Press, 2006.
- [242] J. Yates, P. Griffin, L. Hood and J. Zhou. Computer aided interpretation of low energy MS/MS mass spectra of peptides. In J. Villafranca, editor, *Techniques in Protein Chemistry II*, pages 477–485. Academic Press, San Diego, 1991.
- [243] J. A. Yergey. A general approach to calculating isotopic distributions for mass spectrometry. *Int. J. Mass Spectrom. Ion Phys.*, 52(2–3):337–349, 1983.
- [244] J. Zaia. Mass spectrometry of oligosaccharides. *Mass Spectrom. Rev.*, 23(3):161–227, 2004.
- [245] J. Zhang, E. Gonzalez, T. Hestilow, W. Haskins and Y. Huang. Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Curr. Genomics*, 10(6):388–401, 2009.
- [246] J. Zhang, D. Xu, W. Gao, G. Lin and S. He. Isotope pattern vector based tandem mass spectral data calibration for improved peptide and protein identification. *Rapid Commun. Mass Spectrom.*, 23(21):3448–3456, 2009.
- [247] N. Zhang, R. Aebersold and B. Schwikowski. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, 2(10):1406–1412, 2002.
- [248] W. Zhang and B. T. Chait. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.*, 72(11):2482–2489, 2000.
- [249] R. Zubarev and M. Mann. On the proper use of mass accuracy in proteomics. *Mol. Cell. Proteomics.*, 6(3):377–381, 2007.