Chapter 2

Biological Mass spectrometry

2.1 Mass spectrometry

In 1906 J. J. Thomson was awarded the Nobel Prize in physics for the investigations on the conduction of electricity by gases. Six years later, in 1912, J.J. Thomson reported about the possibility to separate molecules based on their differences in mass and charge. The instrument constructed by Thomson was called parabola spectrograph. The ions generated in the discharge tubes were passed into an electromagnetic field, which made the ions move through parabolic trajectories. The ions were then detected on a photographic plate or fluorescent screen.

Mass Spectrometers used today have the same subunits than the first mass spectrometer of J. J. Thomson. They are assembled of an ion source, mass analyser, detector and a processing unit (Figure 2.1). In the ion source the analyte is ionised and transferred into gas phase. Different types of ion sources were developed in order to ionise molecules of various physical and chemical properties *e.g.* electron ionisation and chemical ionisation for the analysis of liquids and gases or inductively coupled plasma sources used for the analysis of metals. Ionisation techniques are crucial to determine what types of samples can be analysed by mass spectrometry. The ions are then introduced into the mass analyser. The mass analyser is a region of high vacuum in which the ions can be separated according to their mass and charge by employing static or oscillatory electric, magnetic or electromagnetic fields. A large variety of mass analysers

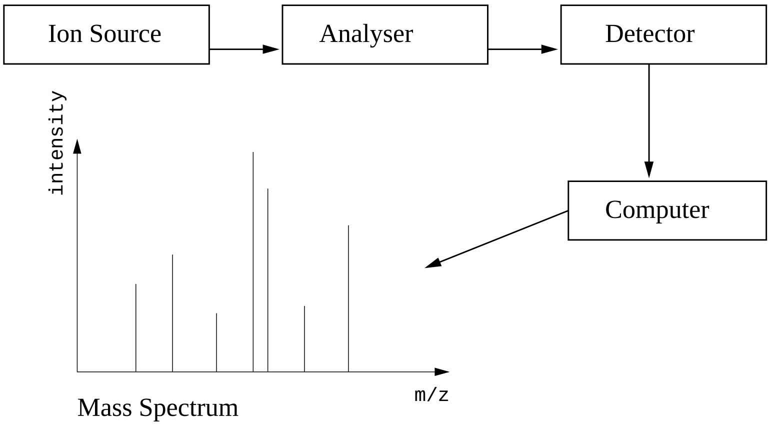


Figure 2.1: Basic components of a spectrometric instrument: ion source, mass analyser, detector and processing unit.

are in use. These include ones, which employ static electric (time of flight), static magnetic or magnetic (sector instruments), oscillating electric (quadrupole, orbitrap), or oscillating magnetic (Fourier Transform ion cyclotron resonance) forces to separate the ions of the analyte according to the mass and charge. The detector records the ions leaving the mass analyser. Typically some types of electron multipliers are used. Because the number of ions leaving the mass analyser at a particular instant is typically quite small, significant amplification is often necessary to get a signal. Microchannel Plate Detectors are commonly used in modern commercial instruments. In FTMS and Orbitraps, the detector consists of a pair of metal surfaces within the mass analyser/ion trap region, which the ions only pass near as they oscillate. Finally, the processing unit integrates the ions recorded by the detector. A schema of a mass spectrometric instrument is shown in Figure 2.1.

2.2 Protein mass spectrometry

Biological mass spectrometry aims to measure the masses of biological molecules. The main difficulty to overcome in order to measure proteins and peptides is to transfer them from the aqueous phase into the gas phase and to ionise them. Proteins are polar molecules consisting of polar and non-polar amino acid residues joined by peptide bonds. The energy required to volatilise molecules in the ion source is commonly supplied as heat. Molecules containing polar groups can form very strong hydrogen bonds with other molecules and are not susceptible to volatilisation by heating. Therefore, in order to enable heat volatilisation chemical modification of the polar groups is necessary.

For the first time in 1966 volatilised thermolabile biomolecules were ionised using gas ions, a method called *Chemical Ionisation*, as described by M.S.B. Munson and F.H. Field (63)). Ten years later (1976) MacFarlane et al. (64) introduced the *Plasma Desorption* ionisation method where high-energy ions were used to desorb and ionise the analyte molecules. Using this technique it was only possible to ionise molecules of masses less than $10kDa$. However, the molecular weight of common proteins ranges from a few thousand Dalton (Da) (alternatively called unified atomic mass unit u) for the hormone insulin ($5734Da$) up to several million Da in case of the muscle protein titin with the molecular weight of $\approx 3000000Da$. Therefore, protein masses are measured in kilo Daltons (kDa).

The Fast atom bombardment (FAB) method introduced by M. Barber et al. (65) enables to ionise polar and thermally labile compounds with size less than $10kDa$. Fast atom bombardment (FAB) is an ionisation technique used in mass spectrometry in which an analyte and a non-volatile chemical protection environment (liquid matrix) mixture are bombarded by $\approx 8KeV$ particle beam of usually inert gas such as argon or xenon. Common matrices include glycerol and 3-nitrobenzyl alcohol (3-NBA). Despite its limitations, FAB provided the foundation of desorption and ionisation methods applied today to large biomolecules: *Electrospray Ionisation* (ESI) and Matrix Assisted Laser Desorption Ionisation. In the FAB technique an analyte and a non-volatile chemical protection environment (liquid matrix) mixture is bombarded by $\sim 8KeV$ particle beam of inert gas such as argon or xenon. This technique is very closely related to liquid matrix secondary ion mass spectrometry (LSIMS). FAB is a relatively soft ionisation technique and produces primarily protonated or deprotonated molecules.

2.3 Peptide Mass Fingerprinting by Matrix-Assisted Laser Desorption Time of Flight Mass Spectrometry

Introduction Masses of the peptides obtained by a mass spectrometric measurement of a sequence specific protein digest are used to identify proteins by comparing them with peptide masses predicted from protein sequences (18; 27). It is expected that the peptide masses of a sequence specific protein digest are uniquely identifying the protein. This method of protein identification is called peptide mass fingerprinting in analogy to a method of identifying people based on the uniqueness of papillary lines on their fingertips. Figure 2.2 illustrates schematically the principle of protein identification by PMF. The purified protein is digested employing a sequence specific protease (Figure 2.2, top right). The protein molecule is cleaved at the specific cleavage sites, determined by the cleavage pattern of the protease *e.g.* Trypsin only cleaves after the amino acid residues KR but not if followed by P and hence, in case of complete cleavage, producing every time the same set of peptides. The mass of each peptide is the sum of the amino acids in the peptide chain including any amino acids modifications. The masses of the peptides can be determined with high precision employing a mass spectrometric instrument. The peptide mass fingerprint is then compared to theoretical peptide masses computed from the sequence specific *in silico* protein digests (Figure 2.2, left).

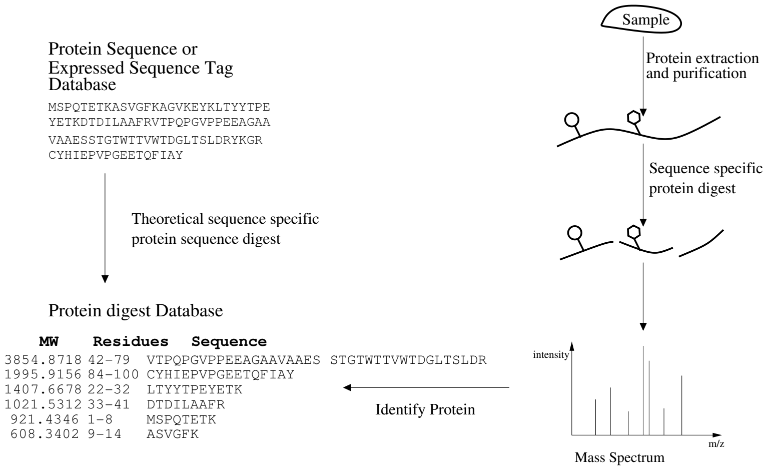
In order to obtain a reliable identification of the protein, the set of peptide masses must be derived from a single protein only. If the sample is contaminated with peptides of other proteins the correct assignment to sequence database entry becomes very difficult. Usually, the researchers are interested in analysing the protein content of the cells, cell components or protein complexes containing up to thousands of different proteins with concentrations varying over many orders of magnitude. Therefore, in order to identify the proteins by PMF it is necessary to separate them first. The protein separation method of choice is the Two Dimensional- Polyacrylamide Gel Electrophoresis ($2D - PAGE$) (66), which enables to separate hundreds of proteins according to their size (number of amino acid residues) and isoelectric point.

The proteins are first separated according to their isoelectric point in a long strip of a gel with a pH gradient. The protein mixture is added at the centre of the gel strip and an electric field of several kV is applied. The proteins migrate through the gel until the pH of the protein equals the pH of the gradient. At this specific point the number of protonated and deprotonated polar groups equals, the charge of the proteins becomes zero and the proteins stop to move in the electric field. This point is called an isoelectric point and the separation method is called an isoelectric focusing.

Next, the gel strip with the proteins separated according their isoelectric point, is soaked in a solution containing high concentrations of sodium dodecylsulphate (*SDS*). SDS is an ionic surfactant that causes the proteins in the gel strip to unfold (denature). The SDS molecule has an acidic sulphate group, which is deprotonated giving the SDS molecule a negative charge. SDS molecules strongly bind to the protein backbone. The gel strip is then aligned on one side of a large and very flat rectangular gel soaked with SDS, followed by the application of an electric field across the gel. The electric field is directed orthogonal to the first separation direction. The separation occurs now according to the length of the denatured protein, proportional to the length of the protein backbone. Small proteins move faster, while large proteins move slower through the mesh of the PAGE gel. The successive application of both techniques in perpendicular directions (two dimensions) enables to separate hundreds of proteins.

In order to reveal the position of the proteins in the SDS-PAGE gel, it is usually stained with colloidal Coomassie blue, silver (gold) or fluorescent dyes, *e.g.* SYPRO Ruby Protein gel stain, compatible with further mass spectrometric measurements. The proteins become visible as spots. The stained spots, which ideally contain a single protein, are than detected and excised. The intensity of the staining can be used for comparative *2D-page* – a method to detect changes in protein concentrations among related samples. For example, it is possible to look for proteins which expression varies in response to diseases (67), for disease diagnostics (68) or to study differences in protein expression among closely related organisms (69).

The excised gel pieces are afterwards washed to remove the SDS and soaked in a solution containing a proteolytic substance. Frequently, the proteolytic substance itself is a protein and hence susceptible to auto-proteolysis. To avoid contamination of the sample peptides with protease peptides biotechnologically modified proteases robust to autoproteolysis are used, and the experimental settings are optimised such that low concentrations of protease have to be used. The proteolysis is typically carried out overnight. The peptides are afterwards eluted from the gel with acetonitrile, dried under vacuum and finally dissolved in a small amount of distilled water. Thus obtained peptide solution can be then analysed by PMF.



```
MSPQTETKASVGFKAGVKEYKLTYYTPE
YETKDTDILAAFRVTPQPGVPPEEAGAA
VAAESSTGTWTTVWTDGLTSLDRYKGR
CYHIEPVPGEETQFIAY
```

```
   MW     Residues   Sequence
3854.8718 42-79   VTPQPGVPPEEAGAAVAAES STGTWTTVWTDGLTSLDR
1995.9156 84-100  CYHIEPVPGEETQFIAY
1407.6678 22-32   LTYYTPEYETK
1021.5312 33-41   DTDILAAFR
 921.4346 1-8     MSPQTETK
 608.3402 9-14    ASVGFK
```

Figure 2.2: Schema of a Protein Mass Fingerprinting (PMF). Given a database of protein sequences (top left) a sequence specific *in silico* database digest is computed (bottom left) and theoretical spectra are modelled. These theoretical spectra are compared with experimental spectra (bottom right). Top left – and experimental protein sample is first separated (not shown). Afterwards, the isolated protein is digested using a sequence specific protease (left). The masses of the peptides are determined by a mass spectrometric measurement to obtain a PMF (bottom left).

Mass spectrometric measurement Usually, one micro-litre of the peptides dissolved in distilled water are applied onto the MALDI target (mass spectrometric sample support – frequently made of stainless steel) and mixed with a UV absorbing matrix. A good mass spectrometric matrix should possess the following properties: it must be soluble in a solution with an analyte, should have a strong absorption coefficient in solid state at the wavelength of the laser and finally be chemically inert in terms of reactivity with the analyte. Mass spectrometric matrices typically used for matrix assisted laser desorption ionisation are: 3,5-Dimethoxy-4-hydroxycinnamic acid (*synapic acid*), alpha-Cyano-4-hydroxycinnamic acid (*alpha cyano*) (70) and 2,5-dihydroxybenzoic acid (*DHB*) (52).

While the water evaporates, matrix and peptide molecules co-crystallise. Afterwards, the mass spectrometric target is inserted into the vacuum chamber of the mass spectrometer. The matrix molecules absorb the high energy light of the laser pulse, explosively evaporate - desorb, and volatilise the sample (Figure 2.3). Furthermore, the matrix molecules are acidic and donate positively charged protons to the peptides. MALDI-MS analysis of small peptides ($< 5kDa$) usually produces only single charged molecular ions. During the ionisation peptide molecules are competing for protons and some peptides are ionised at the expense of others.

Peptides bearing amino acids with proton accepting basic groups in the chains (K, R, H) have better chances to be ionised and therefore detected. Other peptides can be suppressed because of the unfavourable chemical property of $i.e.$ many acidic groups (D, Q). Because of this peak suppression effect, high sample concentrations may be detrimental for the quality of the peptide mass spectrum. Furthermore, the ionisation properties of the peptides make direct quantification of peptides or protein abundance by MALDI-MS impossible. However, relative protein quantification using MALDI-MS is possible if two proteins of different mass but identical chemical properties obtained by isotope labelling are analysed within a single mass spectrometric measurement and compared (71).

Time-of-Flight mass analyser. The Time-of-Flight mass analyser (Figure 2.4) is frequently used with the MALDI ion source. After the laser pulse has volatilised the sample the ions are accelerated in the acceleration region by applying an electric field of several kV (Figure 2.4, left). The acceleration region has a size of several millimetres. The ions reach a constant kinetic energy of

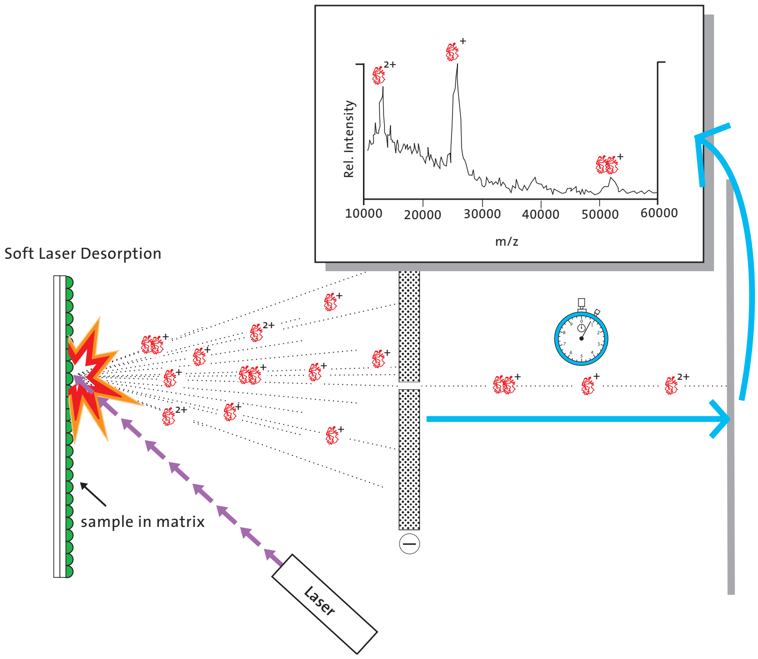


Figure 2.3: Soft Laser Desorption Process (1)

$\sim 25keV$. The velocity is inversely proportional to the square root of the charge over mass m/z. Thus in case of MALDI the peptide charge is usually +1, and in practise it is inversely proportional to the square of the mass. Therefore, the ions can be separated according to their mass m in the field free drift region (Figure 2.4). The m/z of an unknown ion can be determined by comparing its time of flight with the time-of-flight of a known standard. To describe a time of flight mass analyser no more than Newtonian physics is required (58). The set of the measured peptide masses, obtained from a single protein, is called the peptide mass fingerprint.

Summary Figure 2.5 summarises the types of experimental errors introduced at each step of the PMF analysis. For example, during protein extraction, separation and purification of the protein samples contaminants can be introduced. Typical contaminants are keratins - structural proteins of human hair and skin. During the protein digestion, which is usually performed using a sequence specific protease the sample is contaminated by protease peptides resulting from

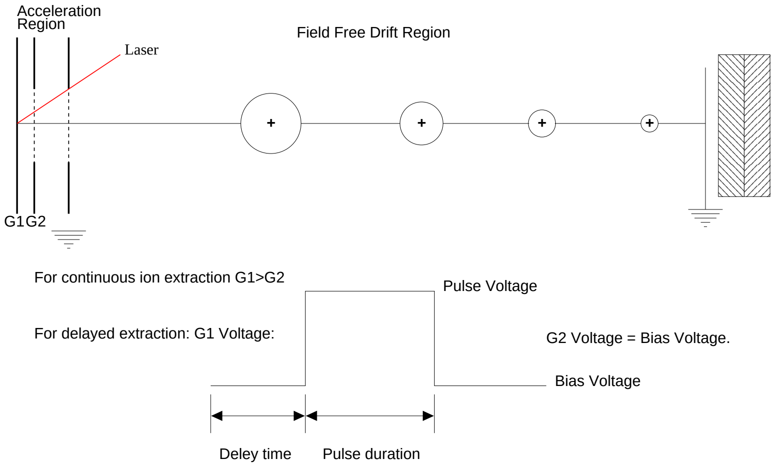


Figure 2.4: Schema of a Time of Flight Spectrometer.

autoproteolysis.

The laser pulse volatilises not only the peptide molecules but also the mass spectrometric matrix molecules. Hence the time of flight of both peptides and matrix is measured. Volatilised matrix molecules usually do not interfere with the peptide mass measurement. Matrix peaks have a much lower mass as compared to the sample peptides, what makes it easy to separate matrix peaks from peptide mass peaks. However, matrix clusters can also be formed during the desorption process and generate signals in the same mass region than peptides (Figure 2.5, bottom right), thus contaminating the peptide spectra. Finally, due to the limited measurement accuracy of the mass spectrometric instrument (41), properties of the detector (3), systematic as well as independently and identically distributed mass measurement errors are introduced (Figure 2.5, bottom left).

2.4 Protein identification by Electrospray ionisation tandem mass spectrometry

Figure 2.6 illustrates schematically the protein identification by tandem mass spectrometry (MS/MS). In case of this protein identification method, the 2D gel electrophoretic separation can be omitted. The complex protein mixture is digested with a sequence specific proteases and a peptide mixture of high

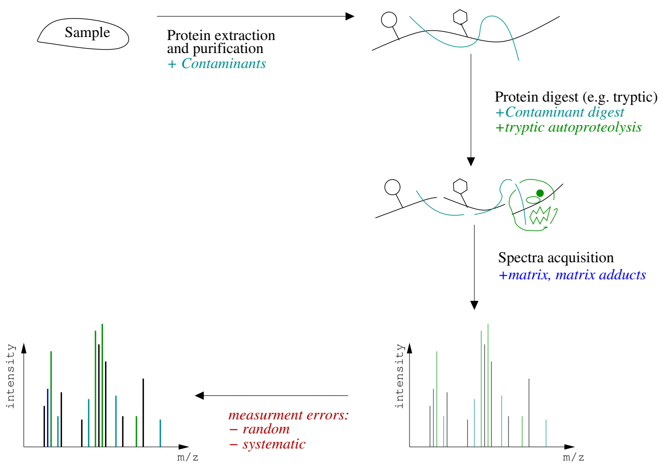


Figure 2.5: Problems of Protein identification by MALDI-TOF. Top – protein contamination by *e.g.* keratins can be introduced during the protein separation. Top left – Autoproteolysis products of the protease, contaminating the sample. Bottom left – During the desorption process, matrix molecules are volatilised, ionised and detected allowing the contamination of the spectra. Because of the limited mass measurement accuracy of the MS instrument, a mass measurement error is introduced (bottom).

complexity is generated. In order to measure and identify peptides in this mixture, the complexity of the sample introduced into the mass spectrometric instrument must be substantially reduced, usually by applying one or multi-dimensional high performance liquid chromatography.

Peptide separation by high performance liquid chromatography In high performance liquid chromatography the analyte is preloaded on a column (stationary phase), which is either polar (Normal Phase Chromatography) or lipophilic (Reversed phase high performance liquid chromatography) and is then eluted from it by a solvent. The solvent (mobile phase) is applied under high pressure as a time dependent gradient of water (polar) and less polar organic liquids *e.g.* methanol or acetonitrile. Less polar peptides elute under conditions of high percentage of ethanol (less polar eluent), while polar peptides (hydrophilic) elute when the water content (a polar eluent) of the mobile phase is high. In order to obtain narrow peaks (better resolution) the volume of the column as well as the elution time needs to be small. The high pressure of the eluent applied onto the column reduces the elution time minimising peak widening due to diffusion.

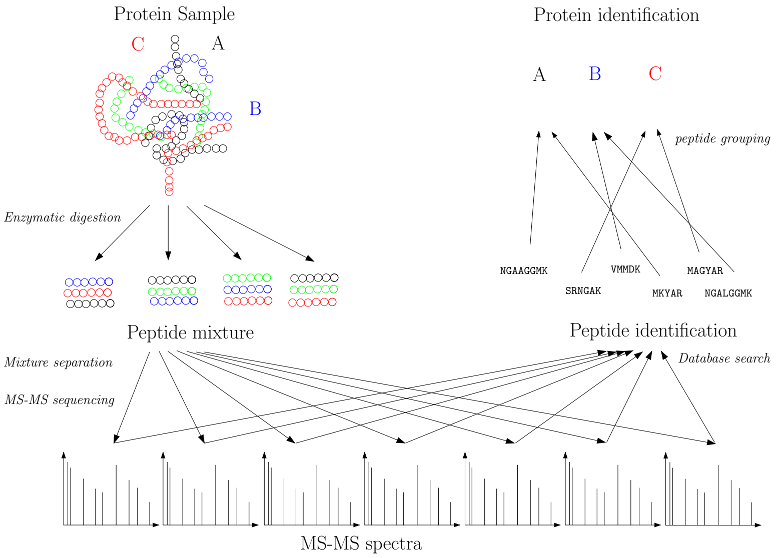


Figure 2.6: Schema of protein identification using tandem mass spectrometry. The protein sample (top left) contains proteins (A, B, C). The protein mixture is digested using sequence specific proteases. The *peptide mixture* is separated using liquid chromatography (LC) technique and MS/MS spectra of peptides are acquired. The peptides are identified by database searches (bottom right). By applying peptide grouping and data analysis (top right) the proteins A, B, C are identified (2).

Electrospray Ionisation Mass Spectrometry HPLC separation generates a constant flow of the peptides eluted from the chromatographic column, which can be directly coupled with the electrospray ionisation ion source. The analyte solution is passed through a needle held at high potential (Figure 2.7, left) and charge is transmitted onto the solution. These highly charged droplets evaporate rapidly in an atmosphere of a protective gas. The number of repulsive electrostatic charges on the droplet surface (surface charge density) increases rapidly until the repulsive forces cause a *Rayleigh explosion* of the droplet. This leads to a number of charged smaller droplets, having in sum a larger surface and smaller surface charge density than the parent droplet. Finally, a spray cone of fine droplets forms at the tip of the needle. Subsequent evaporation and explosions finally lead to molecules of ionised analyte. Peptide molecules usually have charges of

$z = +1, +2, +3$. Malcolm Dole described the process of the analyte electrospray ionisation in the charge residue model published already in 1968 (72), which is still valid today.

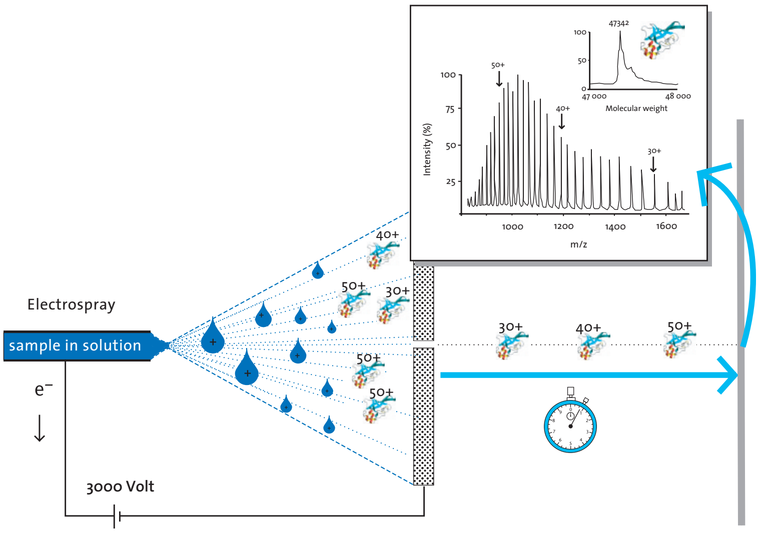


Figure 2.7: The electrospray process (1).

John Fenn was awarded the Nobel Prise in Chemistry 2002 for using the Electrospray process in order to ionise biological macromolecules. He improved Doles's method by using a counter-flow of inert gas in order to aid volatilisation, and combined this ionisation technique with mass spectrometric measurements in order to identify polypeptides (73). The research group of Aleksandrov developed independently at the same time similar methods (74).

The electrospray ion source is usually coupled with a tandem mass spectrometer. Tandem mass spectrometry involves multiple steps of mass selection and analysis separated by fragmentation. In the first mass measurement an intact peptide ion called *parent ion* (Figure 2.8) is selected. Afterwards, the selected peptide ion is fragmented due to collisions with inert gas molecules such as helium, nitrogen or argon (75). The collision converts a part of the translational energy into internal energy of the ion resulting in fragmentation.

The fragmentation patterns of peptides depend on the collision energy. At higher energies, the fragmentation of the amino acid side chains is observed (75;

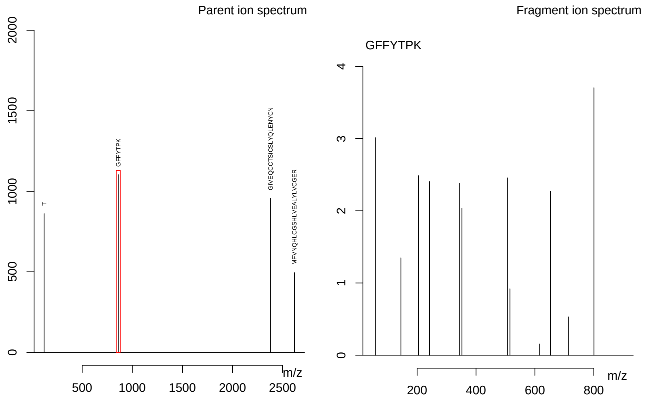


Figure 2.8: Principle of tandem mass spectrometry. Top panel: Parent ion spectrum presenting peptide mass peaks of tryptic digest of the protein hormone insulin. Bottom panel: Fragment ion spectrum of the *parent peptide* mass enclosed in the red square in the top panel with peptide sequence GFFYPTK.

76), while at low energies ($< 100eV$) the fragmentation occurs almost exclusively along the peptide backbone bonds. The peptide backbone break produces a N and a C terminal peptide fragments (Figure 2.9). Depending on the position of the break the N terminal fragments are denoted a, b and c while the C-terminal fragments are denoted x, y and z. The dominant ion series produced by *collision-induced dissociation* (CID) are y and b ions. The fragmentation products are subsequently analysed by a second mass spectrometric measurement (Figure 2.8). The mass peaks of the y and b ion series for the peptide `GFFYPK` are presented in the MS/MS fragment ion spectrum Figure 2.8, bottom panel. Low energy gas phase collision induced dissociation is a preferred fragmentation method for peptide sequencing by mass spectrometry.

Low energy CID MS/MS spectra are typically acquired using ion trap instruments or quadrupole time of flight. Hans Dehmelt and Wolfgang Paul developed the ion trap instrument, and were awarded the Nobel Prize in Physics in 1998 for ”contributions of importance for the development of atomic precision spectroscopy”. In both types of instruments (Quadruple time of flight and ion trap) resonance frequencies of charged particles in an electromagnetic field are exploited. The probability to observe an ion has a maximum at its *resonant*

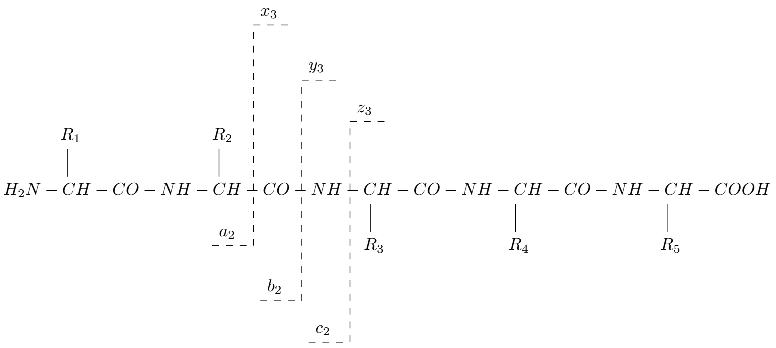


Figure 2.9: CID induced fragmentation pattern and Fragment ion nomenclature. N-terminal a_2, b_2, c_2 ions and C-terminal x_3,y_3,z_3 ions for a five amino acid peptide.

frequency. The resonant frequency is a function of the mass and charge of the ion.

Problems of peptide identification by tandem mass spectrometry Two major obstacles can be observed while identifying the peptides in high throughput experiments by tandem mass spectrometry. The first is the redundancy of peptide identification. Samples for the mass spectrometric measurements are acquired from the continuous flow of the HPLC eluent. Even if the peptide peaks are sharp, the peptide content of many fractions can be almost identical if the sampling rate of the mass spectrometric instrument is high. In order to suppress the acquisition of many identical spectra modern MS/MS spectrometric instruments are able to exclude parent ions masses already measured from repeated fragmentation for several minutes. Nevertheless, redundant measurement and identification of the same peptides in large datasets is observed (Chapter 5). Secondly, the parent ions for subsequent fragmentation are selected, according to their intensity in the primary MS spectrum. Only peaks of high intensity guarantee sufficient sample abundance for subsequent fragmentation, and are selected for fragmentation while MS/MS spectra of peptides with low abundance cannot be acquired. In general MS/MS analysis requires a larger amount of protein sample than identification by PMF.

The second obstacle is that only a relatively low proportion of acquired MS/MS spectra can be assigned to peptide sequences by database searches.

This might be due to the low quality of the peptide ion fragment spectra or inaccurate modelling of the peptide fragmentation patterns. For example, if a modification present in the peptide analysed is not considered when computing the fragmentation patterns some of the search algorithms will fail to assign the spectrum to a sequence. The same problem will occur in case of a single unknown amino acid substitution in the peptide sequence. In this cases *de novo* sequencing methods, or partial sequencing, are of particular interest.

2.5 Protein identification using Mass spectrometric data and sequence information

Protein identification using mass spectrometric data is either based on the comparison of experimentally determined with theoretically predicted peptide masses, or on the analysis of experimentally determined peptide masses in order to derive the protein sequence (i.e. *de-novo* sequencing). The theoretically predicted peptide masses are derived from protein sequences utilizing information about fragmentation patterns. Having set the theoretical spectra representation of all protein- (PMF) or peptide- (MS/MS) sequences in a sequence database, the aim of PMF or MS/MS search is to identify the sequences which most likely generated the experimental spectrum, and furthermore to provide a score which allows the application scientist to asses the significance of the assignments made. The score expresses information such as the number of matching peaks, mass measurement accuracy, number of masses allocated to unmodified and modified peptides, number of not matching peaks, agreement of observed and predicted relative peak intensities and database size. Scoring schemes differ with respect to which kind of data in addition to the peptide masses they incorporate, and which mathematical transformation they use to combined and weight them.

2.5.1 Feature extraction

In order to assign a mass spectrum to a protein sequence first the masses and intensities of peptide or peptide fragment peaks must be determined in the spectrum. This data processing step is usually called *peak picking*. In order to illustrate this process we will describe a two-tiered peak picking method proposed by Lange et al. (3) and Kreitler (77).

In the first step a decomposition of the raw spectrum using signal theoretic methods is applied. The spectrum is decomposed into the low frequency baseline, the analytical signal carrying the significant information (78) and the high frequency noise. This separation can be achieved using signal-processing techniques such as Fourier Transform (77; 79) or wavelet transform (3; 79). Figure 2.10 envisages a wavelet decomposition of an experimental mass spectrum (panel A) into a low-frequency baseline (panel B), an analytical signal (panel C), and into a high frequency noise term (panel D). The analytical signal (panel C) can be used to determine very fast and reliable the approximate location of the peptide peaks (Figure 2.10, panel C).

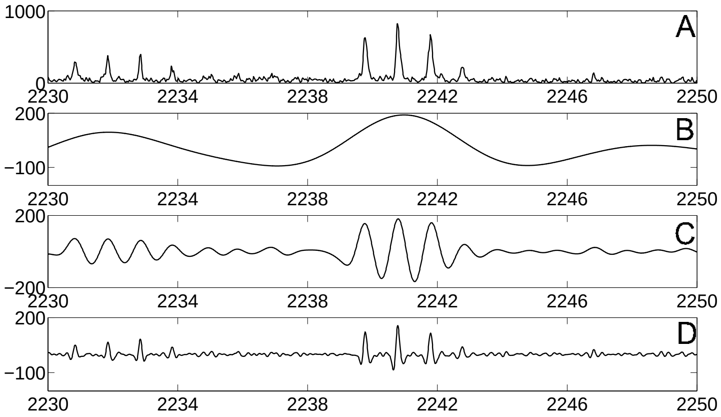


Figure 2.10: In plot A, B, C, and D the x-axis represents the mass interval between 2230Da and 2250Da, whereas the y-axis shows the intensity. A: Part of a MALDI mass spectrum. Plots B, C, and D show the continuous wavelet transform of the spectrum using a Marr wavelet with different dilation values a (B: a = 3, C: a = 0.3, D: a = 0.06) (3).

However, this decomposition might remove information from the analytical spectrum required to determine peak location with high precision. Therefore, in a second step an analytically given asymmetric peak function (3) is fitted to the raw spectrum in the vicinity of the previously determined approximate peak locations. More approaches to the peak-picking problem in proteomics data have been proposed (3; 80; 81; 82; 83; 84), but are not discussed here.

Subsequent steps of spectra pre-processing prior to database search include *calibration* (41), removal of non-peptide peaks (32), peak-list clustering (50) and are described and discussed in greater detail in chapter 3,4 and 5, respectively.

2.5.2 Database scoring schemes

2.5.2.1 Peptide Mass Fingerprinting

MOWSE The MOWSE scoring scheme was one of the first scores to incorporate the frequency of peptides in the database (27). Given a protein sequence of infinite length the frequency of peptides of length N-residues is proportional to $f_C(1 - f_C)^{N-1}$, where f_C is the fractional abundance of the cleavage sites. Protein sequences have however finite length, which influences the frequency of the peptides. To take this into account the frequency of the peptides in the database is computed according to the length (mass) of the *parent proteins*.

Before computing the score the algorithm processes the sequence database and computes the matrix M of peptide frequencies $f_{m_{prot},m_{pep}} = f(x_i < m_{prot} < x_{i=1}, y_j < m_{pep} < y_{j+1})$. In plain English $f_{m_{prot},m_{pep}}$ is the frequency of peptides (m_{pep}), in a mass window $(x_i, x_{i+1}]$, which were generated by cleaving all proteins of mass $y_j < m_{prot} < y_{j+1}$.

Then the MOWSE score is defined by:

$$score = \frac{50000}{m_{prot} \times \prod f_{m_{prot},m_{pep}}} \quad ,$$

where $\prod f(m_{prot}, m_{pep})$ is the product of frequencies of the matching peaks. If the protein frequencies $f(m_{prot}, m_{pep})$ are small the score gets larger. The mass of the protein in the denominator m_{prot} is a normalisation term accounting for the fact that large proteins generate on average more matching peptides. Larger MOWSE scores indicate more significant assignments.

The Probability Based Mascot Score PBMS (55) extends the MOWSE score by incorporating in addition the number of not matching peaks and the mass measurement accuracy. Furthermore, it casts the score into a probabilistic framework. Despite the fact that the exact scoring model for the PBMS score was never disclosed, this scoring schema implemented in the proprietary Mascot search software (4) is among the most widely used. The PBMS represents the probability P that the observed match is a random event and is defined as $-10\log_{10}(P)$.

Wool Smilansky The scoring schema, which considers similar parameters than the PBMS score and is similarly interpretable, is the scoring schema proposed by Wool and Smilansky (30). The authors defined the scoring function F_s with parameters: a) list of monoisotopic peak masses $m1, m2, \ldots m_m$ b) list of theoretical peptide masses $p1, p2, \ldots p_n$ of the protein candidate, c) the measurement accuracy acc, d) the protein database DB, with N_{pept} peptides. From parameters (a, b, c) the number of matches between theoretical and experimental peak masses k can be computed.

The score F_s is defined by the probability that selecting randomly and independently n peptide masses from the database (according to the mass distribution in this database) will generate k hits with m experimental peptide masses. Using the binomial distribution, the probability for k random matches in a protein with n peptides is given by:

$$F_s = Pr(n, k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Similarly to the probability P computed by the Mascot search engine this score is considered good if it is very close to zero, meaning a very small probability of being a random match.

To illustrate their scoring schema the authors give the following example. Assume, a peak-list with $m = 50$ peaks; a protein candidate with $n = 100$ peptides; a measurement accuracy of $acc = \pm 0.05Da$; eight $k = 8$ matches; and a database that contains $N = 20000000$ peptides. Furthermore, we have determined that on average there are $B = 100.000$ peptides in a mass window of size acc of the sequence specific database digest. Therefore, the probability of sampling independently any of these peaks is $p = 100.000/N = .005$. Then the F_s score evaluates to:

$$\binom{100}{8} \times .005^8 \times 0.995^{92} = 4.58 \cdot 10^{-8} \; .$$

The significance of the F_s score depends on the number of proteins N in the database. We expect on average $N \cdot F_s$ protein entries to receive a score of F_s or less. The number $N \cdot F_s$ is often called the $E - score$, as in several well-known sequence search algorithms like $BLAST$ (85). The E-score in case of the example given above will be $600000 \times 4.58 \cdot 10^{-8} = .02748$.

A further prominent and widely used algorithm for scoring of PMF spectra is the Profound score proposed by Zhang and Chait (29).

2.5.2.2 MS/MS peptide ion fragmentation pattern search

Protein identification by searching sequences databases using MS/MS peptide fragment spectra requires two steps of analysis. First, the peptides are assigned to peptide sequences by comparing them with theoretical mass spectra simulated according to the peptide sequences and information about peptide fragmentation patterns (Figure 2.9). Afterwards, the identified peptide sequences are grouped and assigned to protein sequences in order to obtain protein identifications (Figure 2.6, top right). Because this method involves the identification of peptides merely, not only protein sequence databases (most commonly searched) but all types of sequence databases including genomic and *Expressed Sequence Tag* (EST) databases (86; 87) can be searched.

A schematic representation of the spectra identification by peptide ion fragment mass search is shown in Figure 2.11. The experimental spectrum (left) is compared with theoretical spectra (right). The theoretical spectrum represents information about fragmentation patterns of the peptide sequences, ionisation properties of the fragments and mass measurement accuracy.

SEQUEST The SEQUEST (88) scoring algorithm is widely used for the identification of MS/MS spectra. In order to identify a peptide it computes three different types of measures. Two of these measures (S_p and $Xcorr$) express

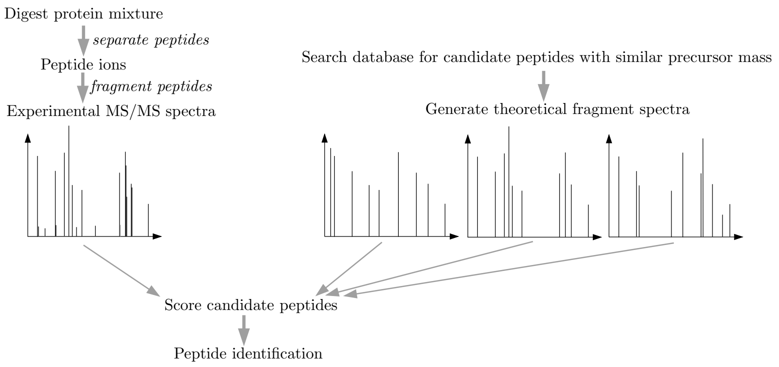

Digest protein mixture
separate peptides
Peptide ions
fragment peptides
Experimental MS/MS spectra
Search database for candidate peptides with similar precursor mass
Generate theoretical fragment spectra
Score candidate peptides
Peptide identification

Figure 2.11: Principle of peptide identification using MS/MS data. The experimental spectrum (left) is compared to theoretical spectra (right). A scoring algorithm compares the this two spectra and computes the significance of the similarity.

the similarity between spectra, while the third (ΔC_n) represents the measure of the significance of the assignment made.

The preliminary score S_p, sums the peak intensity of fragment ions matching theoretically predicted ions, and measures the continuity of an ion series. S_p is called preliminary because it is used to pre-select the candidate sequences for more thoughtful analysis. For all candidate sequences, which possess a sufficiently high S_p score, the theoretical spectra are simulated. The theoretical spectrum includes the b and y fragment ions as well as ions generated due to water and ammonia losses. Because ions generated due to water and ammonia loss are less frequent than the main ion series b and y, a five times weaker intensity is assigned to them as compared to the y and b ions. The mass measurement error is modelled by assigning an intensity of 25% to a mass range of $\pm 0.5 Da$ around the exact theoretical mass of the peptide fragment.

The normalised experimental and theoretical spectra are compared using the cross-correlation measure ($Xcorr$). The cross-correlation is proportional to the number of matching peaks, consequently noisy spectra or spectra derived from long peptides (both have more peaks), will result in higher similarities than for example clean spectra. To account for this, the cross-correlation is normalised by the auto-correlations of the predicted and experimental spectra. The theoretical spectra are then ranked according the normalised $Xcorr$ measure.

The S_p and $Xcorr$ are database independent measures. However, larger databases contain more peptide sequences, and more theoretical spectra will be similar to the experimental spectrum. Therefore, to assess the uniqueness of the assignment, the ΔC_n score that measures the distance between the best and second best match, is computed. The decision if a peptide is identified or not is made by taking into account both the ΔC_n and the $Xcorr$ scores.

SCOPE (89) is a probability based scoring schema for MS/MS data searches. First the probability $\pi(F|p)$ that peptide p gave rise to a fragmentation pattern F is computed. Data mining of curated database of identified MS/MS spectra can derive the probabilities of each cleavage event in the tandem mass spectrum. Furthermore, the probability $\pi(S|F,p)$ that a fragmentation pattern F generated a spectrum S is computed. Into computing $\pi(S|F,p)$ factors such as: the mass measurement accuracy, the agreement of peak intensities, the matching and non-matching of experimental and predicted peaks, are incorporated.

$$\pi(S,p) = \sum \pi(S|F,p)\pi(F,p).$$

The sum must be computed over all possible fragmentation patterns F. To compute the above formula efficiently a dynamic programming algorithm was implemented. SCOPE ranks the peptides p according the score $\pi(S,p)$ and furthermore outputs a corresponding $P - value$.

PeptideSearch (28) correlates the mass differences between mass peaks with amino acid masses to infer a *partial peptide sequence*. Figure 2.12 illustrates the interpretation of MS/MS spectra to derive sequence information. The mass differences in y as well as in b series ions can be correlated with amino acid masses. The *partial sequence tags* can be searched against sequence databases using *Blast* (90) or $FASTA$ style search algorithms. Combining them with spectra comparison approaches described above can increase the specificity and sensitivity of the database searches. Other search algorithms, which combine the inference of short sequence tags from MS-MS spectra with a conventional database searches have been published (38; 91).

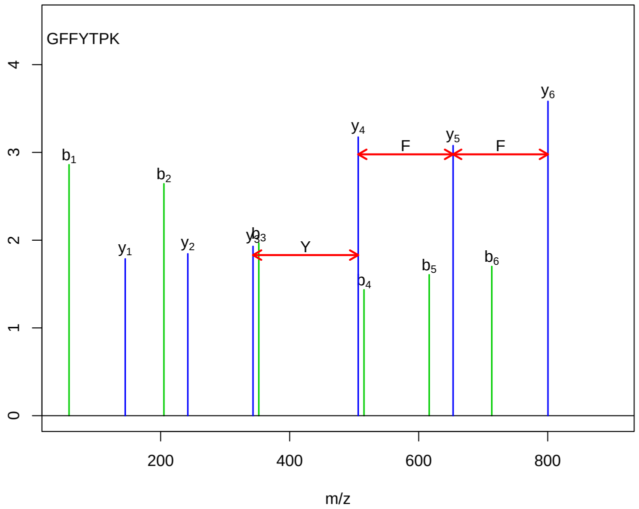


Figure 2.12: Partial interpretation of a spectrum. Simplified representation of an MS/MS spectrum for the peptide GFFYTPK. The b-ion ladder is shown in green, while the y-ion ladder is denoted by blue colour. Distances between peaks on the horizontal mass-to-charge (m/z) axis can be used to infer partial sequences of the peptide. This example shows how the partial sequence FFE can be inferred from the y-ion ladder.

De novo peptide sequencing using MS/MS data. The database search approach described above enables only the identification of peptides that are present in the searched sequence database. It cannot however be used to identify peptides from only partially sequenced genomes or sequence variants of known proteins. Therefore, algorithms for the *de-novo* sequencing of peptide ion fragmentation data were developed (92; 93; 94; 95; 96). The disadvantage of this group of algorithms is that they are computationally very expensive and require high quality MS/MS data. High throughput experiments generate huge amounts of data. Therefore, prior to classification, grouping and merging of experimental spectra must be applied to reduce the amount of data, to increase the quality of the spectra and to make *de-novo* sequencing practically applicable. Hence, in Chapter 5 we studied similarity measures, which can be used to compare and group MS/MS spectra prior to further analysis.

2.5.3 Validation of search results

In addition to the already discussed scores, application scientists are often interested in simple and easily interpretable measures, which help to understand and to validate the search results (97). This easily to interpret measures include the number of matching peaks, the number of matches with peptides bearing modifications in relation to matches with not modified peptides, the distribution of the differences between experimental and theoretical masses or the protein sequence coverage.

A different approach to validate a search result is to use several scoring algorithms and to compare the identification results. When peak-lists are derived from noisy spectra with low mass accuracy and the search is performed in huge sequence databases, different scoring schemes might produce different results. Studies to weight up the performance of scoring algorithms were conducted for example by Chamrad and Meyer (98), however there is not one single best scoring schema. Therefore, several scoring algorithms are combined to validate the correctness of identifications (36).

In case of MS/MS data the task of validating search results can be facilitated by software tools such as INTERACT (99), DTASelect (100) or CHOMPER (101). These programs produce summary statistics of the measures and scores produced by the search algorithms, provide graphical tools to view and compare visually the theoretical and experimental mass-lists, visualise the ion series in the experimental spectra, group the peptides according to the proteins they are derived from, and compute summaries, *e.g.* sequence coverage of the protein (Figure 2.6, top right).

Information, which can be obtained directly from the sequences of assigned peptides, but which is not used by the scoring algorithms themselves, can help to validate search results. For example, the elution time in reverse phase chromatography can be modelled given the peptide sequence using neural networks and compared with the experimentally observed elution times (102). It is also easily possible to compute the isoelectric point from the protein or peptide sequence and compare it with the experimental data (103).

Statistical classification methods can also be employed to determine, based on the scores and measures produced by the search algorithm, if the assignments made are correct. For example Keller et al. (47) applied linear discriminant analysis while Anderson et al (46) used support vector machines. However, in order to train the classifier a dataset where the identities of the spectra are known is required. Various classifiers were used to validate the assignments made by the database scoring algorithms (mainly SEQUEST).

2.6 Summary

To identify proteins both PMF and peptide fragment ion spectra are used. MS/MS data are widely considered to provide more reliable and trustworthy protein identifications. That is because in order to identify a single protein usually many peptide identifications based on ion fragment spectra are required while in case of PMF a single spectrum must suffice to identify a protein (97).

However, due to the high ion transmission of the time of flight mass analyser, the MALDI technique is more sensitive compared with other MS techniques. Furthermore, in relation to Electrospray ionisation MS, MALDI-MS is more tolerant to sample contamination, resulting from salts and detergents often present in protein samples and difficult to remove. Therefore, PMF is still the method of choice if analysing spots excised from 2D-PAGE gels (97).

In order to study the whole proteomes, different bioinformatics techniques necessary for the analysis of large amounts of mass spectrometric and sequence data are equally important as the development of new instrumentation and laboratory techniques. New and fast peak-picking algorithms, automatic, precise and reliable calibration methods, scoring schemes for searches in large databases and algorithms to automatically validate the search results need to be developed (9).