# Chapter 1

# Introduction

In recent years, Mass Spectrometry (MS) has emerged as a powerful technique to identify peptides and proteins in complex biological samples (6; 7; 8; 9; 10). Before the identification of the complex constituents, several separation steps are required to reduce the sample complexity. The classical protein separation method is the two-dimensional gel electrophoresis (11; 12; 13; 14), followed by excision of the detected spots from the gel, digestion with sequence specific proteases and extraction of the generated peptides (15; 16). Mass Spectrometric analysis (9; 10) of the resulting mixture of peptides yields a *peptide mass fingerprint* (PMF): a set of measured molecular masses of the proteolytic peptides derived from the analysed protein (17; 18; 19). Alternatively the sequence specific protein digest can be made prior to separation. The peptide mixture can be separated using chromatographic techniques like capilary electrophoresis (20) or displacement chromatography (21). In this thesis we will concentrate on data separated by reversed-phase *high performance liquid chromatography*(HPLC) and ion exchange chromatography (22; 23). The separated peptides can be subjected to MS/MS peptide fragment ion analysis (24; 25). Prior to the mass spectrometric measurement the peptides have to be ionized. PMF analysis is commonly conducted by employing *Matrix Assisted Laser Desorption/Ionisation* (MALDI) *time of flight* (TOF) instruments, while MS/MS analysis can performed using ESI ion trap instruments. A PMF and MS/MS spectrum is a highly specific set of peptide molecular masses derived from one isolated protein or peptide, respectively. The combination of efficient protein and peptide separation

techniques with mass spectrometric methods has provided a powerful approach of rapid detection and identification of peptides and proteins in complex biological samples (26).

Before performing database searches, the MS spectra are processed and the most informative features, namely the locations and masses of the monoisotopic peptide peaks are determined. A list of *mass over charge* ($m/z$) values of the monoisotopic peaks, and either the area under or the height of those peaks, are obtained. This set of $m/z$ and intensity value pairs is called *peak-list*.

Peptide peak lists can be used to identify the analysed protein in large protein sequence databases by matching the determined peptide molecular masses to values calculated from the amino acid sequences in the database. In order to indicate the significance of the assignment a score is computed which takes into account the frequencies of the protein and peptide masses in the sequence databases (27; 28; 29; 30; 31; 32). Other properties included in the score concern the different sensitivity of detection for individual peptides, known protein modifications, and/or possible mutations (33; 34; 35), although generally, all popular search scores depend on the *precise* assignment of experimental to theoretical peptide masses. Similarly, MS/MS spectra can be used for protein identification by searching the determined peptide fragment ion masses against the predicted ones. The prediction is based on the available amino acid sequence data and fragmentation characteristics of the employed MS instrumentation (36; 37; 38; 39).

The sensitivity and specificity of the peptide identification using database searches can be increased by several methods. This usually includes: calibration (30; 37; 40; 41), identification of non-peptide peaks (30; 32; 42; 43), identification and removal of low-quality spectra (44; 45), validation of the search results using machine-learning algorithms (46; 47), and the pairwise comparison of the peak-lists (5; 37; 48; 49; 50).

In summary, protein identification using mass spectrometry starts with a laboratory experiment where proteins are isolated, mass spectrometric samples are prepared and a spectrum is acquired. It is followed by a computational analysis of the obtained spectra, which includes spectra processing, calibration,

filtering, database searches and validation of the search results. In this work we present contribution to the *in-silico* part of the protein identification.

**Chapter 3** introduces an analytical model to predict the masses of the peptide cluster centres in an *in silico* protein database digest. Gay et al. (51) studied the theoretical distribution of peptide masses in a theoretical digest of protein sequences and observed that peptide masses form equidistantly spaced clusters. The clustering is caused by the mass properties of atoms and the elemental composition of the amino acids. Intrinsic properties of peptide mass give way to many applications, *e.g.* calibration (30; 35; 41), non-peptide peak filtering (52; 53), spectra comparison (5; 50) and database searches. All these applications require knowing the exact distance between the peptide mass cluster centres.

We have developed an analytical model to predict the masses of the peptide cluster centres taking into account the frequencies of the amino acids in the sequence database (54), the average protein length of the proteins in the database, the cleavage sites of the proteolytic enzyme and the cleavage probability. Based on this model, we introduced a measure of the deviation of peptide masses from the nearest cluster centre, which is a refinement of a measure proposed by Wool and Smilansky (30). We have developed a new peptide peak-list calibration method based on this distance measure. With this calibration method we obtained calibration accuracies surpassing the other methods based on the property of peptide mass clustering (30; 35). We have also used the distance measure to identify and remove non-peptide peaks prior to database searches using the Mascot search engine (55).

**Chapter 4** introduces novel calibration methods designed to calibrate samples acquired by parallel spectra acquisition. In case of MALDI-MS high throughput experiments (56; 57), the samples are placed in a grid on a manoeuvrable sample support. Then the laser rapidly shoots light pulses on the specific spot that was moved into the laser beam position. Hence, for each spot spectra are acquired. If the mass spectrometer used is a TOF machine the measured data is the time of flight of the ions. This measurement has to be transformed into mass over charge (m/z) using a quadratic calibration function where calibration coefficients need

to be determined. Usually this is done once for all spot positions. However, the calibration coefficients to transform the TOF into $m/z$ differ depending on sample position. This is due to deviations in plate flatness, sample topography changing the size of the acceleration region (40; 56), and alterations in the strength of the electric field on the sample support borders which influences the final energy of the ions (58).

We have developed two novel calibration methods for PMF data. Both calibration methods exploit similarities of peak-lists due to closeness in the origin of the analysed samples. The first method combines the computation of dissimilarities (50) between peak-lists with internal calibration (42; 43). The second method employs spatial statistical methods (59) to model systematic changes of the mass measurement error over the MALDI sample support. The major advantage of the presented methods originates from the fact that the MS calibration derives from samples without internal standards or external calibrants positioned on each sample support.

**Chapter 5** presents a study of multiple distance measures for the pairwise comparison of mass spectrometric spectra. Direct peak-list comparison may be advantageous for many applications. For example the sensitivity and specificity of peptide and protein identification can be increased by the pairwise comparison of the peak-lists as part of the "subtractive analysis technique" (5; 37; 48; 49). In our work we have reviewed a large group of dissimilarity measures and examined how these can be extended to include the mass spectrometry specific property of mass measurement accuracy. A new parameter *weight of non-matching peaks* was introduced into the computation of distance measures. We have studied the Euclidean and the Manhattan distance, the covariance, the sum of agreeing intensities and the spectral angle. We have also examined the impact of the intensity scaling on the outcome of intensity-based measures (60; 61). In addition, we have performed a systematic study of various intensity transformations (5) in order to determine the best variance stabilising transformation. Furthermore, we investigated quantitative measures, *i.e.* Huberts $\Gamma$ or the relative mutual information measure (62). The analysis of variance method was used to provide insight into the relevance of various factors influencing the outcome of the

pairwise peak-list comparison. For large MS/MS and PMF data sets the outcome of ANOVA analysis was consistent, providing a strong indication that the results presented here might be valid for many various types of peptide mass measurements.