Algorithms for the analysis of MALDI peptide mass fingerprint spectra for proteomics

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät der
Universität Basel

von

Flavio Monigatti

aus Brusio Graubünden

Basel, 2005

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Torsten Schwede
PD Dr. Hanno Langen
Dr. Paul Jenö

Basel, den 16.11.2004

Prof. Dr. Hans-Jakob Wirz
Dekanin/Dekan

Table of Contents

Abstract

In this study we present a simple algorithm which allows accurate estimates of the similarity between peptide fingerprint mass spectra from matrix assisted laser desorption/ionization (MALDI) spectrometers. The algorithm, which is a combination of mass correlation and intensity rank correlation, was used to cluster similar spectra and to generate consensus spectra from a data store of more than 100,000 spectra. The resulting first spectra library of 1248 unambiguously identified different protein digests was used to search for missed cleavage patterns that have not been reported so far and to shed light on some peptide ionization characteristics. The findings of this study could directly be applied to a peptide mass fingerprint search algorithm to decrease the false positive error rate to <0.25%. Furthermore, the results contribute to the understanding of the peptide ionization process in MALDI experiments.

The reference library of consensus spectra was also used to identify MALDI peptide mass fingerprint spectra by comparison of the experimental spectra with the spectra in the library. We report the potential of this method to achieve an identification rate of almost 100%.

In a second step, the information derived from the clustering of similar spectra was used to match similar spectra content on different two-dimensional polyacrylamid gel electrophoresis (2D-PAGE) gels. This is to our knowledge the first attempt to match different gels on the level of mass spectrometric information.

A newly established method that makes use of the new techniques is compared to a proteomics study carried out employing traditional proteomics strategies.

1. Introduction to proteomics

1.1. The proteome

About ten years ago, the term proteomics was introduced by Marc Williams, a post-doc in Canberra, Australia. Proteomics is a hallmark technology of the post-genome era. Since the antecedent of the proteome is the genome, it is a generic term used to assemble the whole complexity of protein expression in one word. However, unlike the genome, the proteome does not denote a unique and permanent feature of a given organism. It is changing with the state of development, the tissue, or even the environmental conditions under which an organism finds itself. Thus, there are many more forms of proteins expressed in a cell than the number of genes makes us believe.

The importance of proteome investigations is given by the fact that there is only a weak correlation between the abundance of a protein and its level of mRNA transcription[1,2]. Proteins can be present in a cell although the corresponding mRNA is not found. Even if there was a relation between the level of mRNA expressed and protein abundance, the complexity of a proteome is additionally defined through other means of modification. In higher organisms, synthesized proteins can undergo several modification steps starting from tissue specific alternative splicing to the post-translational modifications of proteins. These modifications change the look of the proteome and are of high importance for protein function. The modifications can alter depending on the state of the cells. This complexity not only differentiates the proteome from the genome, it also changes the way of capturing knowledge from it. While study of gene expression on the mRNA level is extremely powerful and useful[3], a number of questions cannot be answered studying that level of regulation. A technology is demanded that is able to cope with the many different gene products, on the level of proteins. This technology is what we call proteomics.

Since the proteome can be regarded as the complement of proteins at a given time point, it is necessary to have a technology at hand that is able to provide a picture of the actual protein content.

1.2. Proteomics technologies

The key to the development of proteomics has been Mass spectrometry (MS). It can be used to identify and, increasingly, quantify large numbers of proteins from complex samples. Mass spectrometers consist of an ion source, a mass analyzer and a detector. There are four main types of mass analyzer currently in use: ion trap, time-of-flight, quadrupole, and Fourier transform ion cyclotron.

Ion traps are often used in combination with electrospray ionization. ESI ionizes the analytes out of a solution and thus, can be coupled to liquid-based separation systems. Time-of-flight (TOF) analyzers measure the mass of intact peptides with high accuracy and resolution and are often used for high-throughput protein identification by peptide mapping (peptide mass fingerprinting). Typically, the soft ionization method is matrix-assisted laser desorption/ionization (MALDI) in combination with delayed ion extraction.

A quite different approach to probing protein activity and function is the protein microarray. The analytical microarray contains an ordered array of protein-specific ligands, typically antibodies, spotted onto a derivatized solid surface. They can be used to monitor differential protein expression, protein profiling and clinical diagnostics. However, progress here is constrained by a lack of comprehensive sets of high-specificity, high-affinity antibodies[4].

1.3. Protein isolation and separation for proteomics

Proteomics cannot be realized without generic, sensitive and selective methods to isolate and separate proteins from cells or tissues. One of the best methods to separate proteins is electrophoresis, invented by Tiselius. For proteomics, special popularity was obtained by a high resolution variant of that technology: two-dimensional polyacrylamid gel electrophoresis (2D-PAGE). In the first dimension, proteins are separated on the basis of their charges, followed by a separation in size in the second dimension. The technique of separating the proteins by charge is called isoelectric focusing. The second dimension separation uses sodium-dodecyl-sulfate poly

acrylamid gel electrophoresis (SDS-PAGE) to separate proteins by size[5-8]. Since manufactured immobilized pH gradients[9] and precast SDS-PAGE gels became commercially available, the 2D-PAGE methodology has become a routine separation technique.

Other separation techniques are serial liquid chromatography and capillary electrophoresis. Chromatographic processes can be defined as separation techniques involving mass-transfer between stationary (solid) and mobile (liquid) phases. The analytes are first dissolved in a solvent, and then forced to flow through a chromatographic column under a high pressure. In the column, the mixture is resolved into its components. The amount of resolution is important, and is dependent upon the extent of interaction between the solute components and the stationary phase. Liquid chromatography is mainly used to pre-separate complex protein mixtures (e.g a cell homogenate) into different fractions of interest. This allows reaching a higher dynamic range of separation than with 2D-PAGE alone. Several labs have successfully explored couplings of LC with 1D-PAGE.

1.4. Protein identification methods

Since up to 2-5000 spots can be visualized on a single 2D gel, proteomics requires techniques to identify and quantify the proteins in a high throughput manner. Usually, the spots of interest are excised from the gel and further processed. Historically, this has been done by the cumbersome techniques of N-terminal sequencing and amino acid analysis. With the advent of high resolution mass spectrometry of peptides, a more appropriate method was introduced to identify proteins from a spot. This technique is called peptide mass fingerprinting. The idea behind this approach is that enzymatically digested proteins from a spot result in an assembly of peptides of which the masses are measured by mass spectrometry. The spectrum derived from the measurement of such a peptide assembly denotes a unique fingerprint that in turn allows the identification by searching the database of virtually digested proteins for a matching combination of peptide masses.

Cleveland et al.[10] proposed in 1977 to use a set of peptide masses obtained by enzymatic cleavage as a unique fingerprint. This fingerprint allows the identification of a protein in database searches. The concept of peptide mass fingerprinting as an alternative to peptide sequencing was then re-introduced by *Henzel* in 1989. It remained unused until 1993 when 5 groups independently presented methods and algorithms to search databases using mass spectrometric data[11-15]. Three of the algorithms use a simple scoring scheme to order the proteins according to the decreasing number of matching peptides. A pre-filter that considers the approximate molecular weight of the intact protein eliminates random matches. Highly modified proteins or proteolytic fragments are not identified, because the method assumes that the protein being analyzed does not greatly differ from the virtual protein in the database. This problem was circumvented by using a sliding window that regards masses only occurring in the current range of the window.

Two other papers presented scoring schemes based on probability, where the probability of a random hit in the database is calculated [12,15].

All approaches have in common that the mass accuracy plays an important role. The tighter the mass tolerance, the more stringent the identification. Today's technical improvements in mass spectrometry led to development of machines that produce highly accurate peak masses while maintaining a high level of sensitivity.

Thus, mass spectrometry and peptide mass fingerprinting have become ideal tools for reliable and fast protein identification. For high throughput operation, methods and techniques are demanded to automate the steps from gel handling and mass spectrometric measurements to the identification by database searches. Nowadays, the gels are analyzed by image software that automatically annotates the spots. Information from this procedure is given on to a gel picker, which automatically excises the gel pieces that have been annotated by the gel image analysis software. Another automated step provides for the washing and digestion of the gel pieces to prepare the proteins for mass spectrometric measurements. In order to achieve the highest possible level of reproducibility, standard operation procedures are developed. Modern data acquisition software allows the mass spectrometer to operate without operator intervention, meaning the peptide fingerprint is scanned until a useful spectrum is achieved. Thus,

most of the spectra measured contain valid data that can be used to identify the proteins. Every spectrum measured is then analyzed and resulting peaks are annotated.
For protein identification, the peak list is compared to peptide masses derived from theoretical enzymatic digestion of the protein sequence in the databases. The combination of possible peptide arrangements yields a match probability that can be judged for significance. Significance can be tested using spectra whose masses have been shifted so that identification of proteins would not be possible. It is important to consider realistic sets of peptides and their possible modifications. In terms of probability, it is not helpful to consider too many possible peptide hits. While including well known missed cleavages increases the probability of a match, the loss in specificity when considering all possible modifications is dramatic. This is due to the 10-fold increased size of the database.
In our lab, most of the spectra are measured by matrix assisted laser desorption ionization time of flight mass spectrometers. A short introduction to this method is given below.

1.4.1. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry

Typical applications of Matrix-assisted laser desorption ionization (MALDI[16]) are the investigation of complex samples and the direct characterization of cellular material. After a two-dimensional gel electrophoresis in order to separate proteins in a multi-component sample, MALDI is the technique of choice for characterization. Among the most attractive properties of MALDI-TOF are its high sensitivity and the possibility of using it as high throughput method. MALDI has been evolved from the rather old technology of laser desorption ionization (LDI). This technology, which is still being used, is based on the approach of air–drying analyte solutions on a metal target. The spots on the metal plate are ionized using an ultraviolet laser pulse. The resulting ions can be detected by e.g. time-of-flight (TOF) mass analyzers. The limitation of LDI is that only light absorbing molecules are accessible, whereas non-absorbing molecules cannot be ionized (or only by extensive fragmentation).

In order to overcome the disadvantages, the analyte molecule is prevented from being directly involved in the energetic process necessary for desorption and ionization. Instead, the energy transfer takes place on an intermediate matrix rather than on the analyte molecule itself. The successful ionization of the previously non-absorbing molecule alanine in conjunction with the high absorbing molecule tryptophan as energy acceptor demonstrated the feasibility of this approach. Alanine as a non-absorbing molecule alone was not accessible by LDI. Tryptophan, a highly absorbing molecule due to its delocalized electron system, acted as an intermediate energy acceptor and both components could be detected. This observation led to the development of MALDI by using special UV-absorbing materials (matrix substances) that have the purpose of accepting and forwarding energy in order to ionize the analyte[17].

Matrix substances provide for the separation of the analyte molecules from each other and allow only small interactions between the analyte molecules and the target substance. They are usually volatile in contrast to the analyte molecules. There are many pathways of ion formation, but usually ion formation happens due to protonation and de-protonation via interactions between energy carrying molecules and neutral molecules. Thus, most probably the first event in MALDI ion formation is the protonation of the matrix, induced by the generation of radicals.

Good MALDI matrices are characterized by high UV spectral absorption, a high level of proton acceptance in order to allow the protonation of co-desorbed material and a good mixing compatibility with the analyte molecule. Useful matrices have been found rather empirically than systematically.

Different techniques are used to prepare the samples for MALDI measurements. The most commonly used is the “dried-droplet” method[18]. Approximately 0.5µl of a matrix-analyte solution are dropped onto a metal target plate are dried before measuring it with the mass spectrometer. The drying is either done slowly at room temperature or quickly on a preheated target plate. Usually the matrix to analyte concentration ratio is around 1:1000 and the matrix is saturated. Slowly drying of the solution leads to the formation of large matrix crystals and a high level of incorporated analyte substance. This process yields an additional purification effect on the analyte, however, a negative effect of drying slowly is the rather high sample inhomogeneity across the spot on the target.

Quickly drying the spot causes the formation small spots, which are homogenous. These crystals lack the purification effect due to its size. Typically, sample amounts necessary for analysis are in the range of a few femtomoles.

The most common type of ion source uses a pulsed nitrogen laser at 337 nm for desorption ionization. This is a balance between the UV absorption of the most useful matrices and the wavelengths-dependent ionization behavior. MALDI is typically coupled to TOF mass analyzers. Reflectron instruments are used to compensate the initial energy spreads and improve the mass resolution and mass accuracy[16].

However, the breakthrough of MALDI TOF came with the broad introduction of a technique called delayed ion extraction[19,20]. The principle of delayed ion extraction (DE) is rather old and was occasionally used in the early days of TOF mass analysis[21]. The primary contribution to mass resolution loss in conventional (i.e., continuous ion extraction) linear MALDI TOF-MS is attributed to a range of flight times of identical m/z ions due to different initial velocities. No compensation is made with continuous ion extraction linear TOF-MS for ions with the same m/z but different initial ion velocities. Improvements in mass resolution can be achieved by utilizing delayed pulsed ion extraction, which can compensate for the initial velocity distribution of the MALDI generated ion packet such that ions having identical m/z values arrive simultaneously at a space focal plane located at the detector. Broadening of the ion velocity distribution due to collisional processes in the ion source can also be minimized by allowing the dense plume of MALDI generated ions/neutrals to dissipate and cool prior to ion drawout from the ion source. This results in narrower ion arrival time distributions and provides better mass resolution when compared to continuous ion extraction.

1.5. Two different proteomics strategies using 2D-PAGE

The two experimental pillars of proteomics technology are protein separation and mass spectrometry. These technologies can be used in a variety of ways. For labs that use 2D–PAGE gels as ultimate protein separation step, two different approaches commonly employed. The first, more common approach is based on the differences found when

comparing and quantifying images derived from two-dimensional polyacrylamid gel electrophoresis of two different samples[22,23]. Only the spots that are found to be differentially expressed on either gel are further processed. These spots are cut out and the underlying protein is enzymatically digested and prepared for mass spectrometric measurement. The resulting peptide mass fingerprint is then identified by database searches[23]. The result of such an analysis is a set of all differentially expressed proteins.

The second approach is also based on the separation of the protein mixtures of a sample by two-dimensional gel electrophoresis. However, the second step is not gel comparison or spot quantification. Instead, differences between two samples are derived by differential identification, completely lacking an in depth gel comparison. Thus, every spot detected by image analysis is picked, washed and enzymatically digested. A mass spectrum is generated and the underlying protein is identified using the peptide mass fingerprinting technology[24,25].

The first approach is widely used and has been shown to be successfully applied in different contexts of proteomic research. Nevertheless, applying this method is not free of problems. The main disadvantage using this strategy is that it is very difficult to accurately compare 2D-PAGE images in order to find differentially expressed proteins. Thus, eventual differences are not observed. Secondly, any gel comparison software is unable to compare more than a few gels in reasonable time, so that this methodology is not applicable for high throughput proteomics.

The second approach is a method less often applied, simply because the capacity to carry out such kind of massive mass spectrometric spot evaluation is not given. While the outcome when comparing different identifications instead of different spots is more robust and accurate in comparison to gel image analysis, its main drawback lays in the fact that only 40 – 60 percent of the spots and their corresponding proteins are identified. However, the gain in information content by identification of a whole set of spots and their proteins is remarkable. Figure 1 shows a summary of the two strategies.

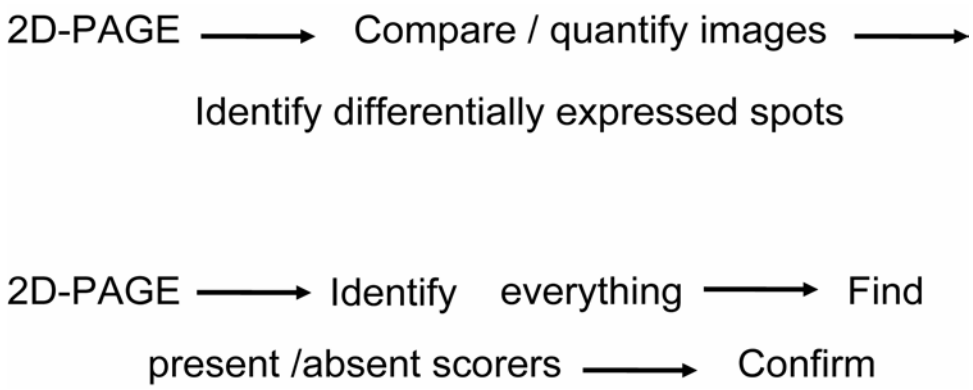


Figure 1: This figure outlines two strategies employed for proteomics studies. Both methods separate protein mixtures using two-dimensional gel electrophoresis. In the first approach, differences in protein content of two samples are quantified by identification of the differentially expressed proteins using gel comparison software. Using the second approach, all spots and proteins are identified and differences are quantified on the level of identifications.

1.6. Application of proteomics

The idea of the proteomics approach is the parallel analysis of expressed proteins at a given time point. As proteomics studies can be conducted at a high level of reproducibility, comparative protein expression pattern analysis can be carried out. Proteomics technology is used to study the influence of toxins or drug treatment on metabolic pathways and the resulting change in protein expression. Quantitative protein expression changes due to exogenous substances can be measured accurately. Recent published works on the use of mass spectrometry as a tool for image analysis shows another field of application. The mass spectrometry laboratory of *Caprioli, R.M.* demonstrated the use of mass spectrometric data derived from scanning tissues of any kind in order to generate a picture of the tissue[26,27]. The spectra measured from the tissue are used to derive a protein pattern for a scanned region on the tissue that allows a histological classification. While previous histological classification methods where based on the staining of only a few peptides, classification based on mass spectrometric data in addition allows a more robust determination of differences when comparing two tissues.

Post-translational modifications of proteins are known to play an important role in the function of biological pathways. Hence, it is desirable to have a technique at hand that

is able to produce results proving a possible modification attached to a protein sequence. Mass spectrometry is a possible tool to detect such modifications[28,29].

2. Aims of the study

As it is outlined in the previous chapter, two different strategies are followed to carry out comparative proteomics studies. However, both technologies have limitations (see chapter 1.3). The aims of this study are the development of a novel methodology in order to overcome the limitations given by the two strategies mentioned in chapter 1.3.

It is crucial for any proteomics study to achieve a maximum identification rate when employing peptide mass fingerprinting. Thus, this study addresses the issues of improving algorithms to obtain better performing peptide mass fingerprinting algorithms. This can be done in two ways. First of all, as the PMF algorithm calculates the probability of a match based on the size of the peptide database, it is crucial to determine a realistic subset of digestion products for the theoretical digest of sequence databases. Thus, the match probability is increased by searching only for peptides that are often observed in mass spectra and neglecting peptides that are rarely or never observed. A second approach to address this problem would be the inclusion of the second dimension of a mass spectrum, namely the peak intensity. So far, no PMF algorithm is known that includes a measure of the intensity of a peak when calculating match probabilities.

As it was mentioned in chapter 1.3 it is very difficult to compare gels accurately using just the images and spot locations. In order to improve the accuracy of gel matching, we are searching for a method to compare gels on the level of mass spectrometric content and not on the level of spot distribution or identification.

The new method we aim to develop is a combination of the two previously employed methods. It should comprise a methodology to carry out comparisons of two-dimensional gel electrophoresis images on the level of the mass spectrometric measurements of the peptides derived from the detected spots. Instead of comparing identifications, comparisons are directly carried out on the spectra itself. Thus, the new schema uses the full information content from MALDI, because the information content of a spectrum is much higher than the information that is abstracted in its identification and much higher than what is contained in an image of a 2D gel.

3. Mass spectrometric data processing

This chapter covers the description of the algorithms and methods we used to measure the spectra and annotate its peaks. It also describes the method employed for searching the protein sequence databases in order to identify the spectra.

3.1. Raw data processing

Spectra were taken from a database of mass spectra of tryptic digests of proteins picked from 2D gels. All protein spots were automatically excised and digested using established protocols[30]. The mass spectrometric measurements were carried out on Bruker Ultraflex instruments (Bruker Daltonics, Bremen, Germany), using ACTH and Bradykinin as internal mass standards. These standard masses have a well defined mass of ~904.5 Da and ~2465.2 Da and are later on used to calibrate the spectra. All spectra were acquired in reflector mode in a mass range between 850 Da and 4200 Da. A spectrum was accepted if after 100–200 scans a minimum peak height and resolution was obtained. This is done using an automated data acquisition method. As explained below, monoisotopic peptide masses were automatically detected from the mass spectra by an in-house peak annotation method. The spectra are filtered for known keratin, trypsin, and matrix fragments. Peak hits are then compared to theoretical masses of peptides derived from an *in-silico* tryptic digest of all proteins from protein sequence databases (e.g. UniProt[31], or NCBI human, mouse, or rat genome draft, as appropriate), in order to identify the protein.

3.2. Peak Annotation

The peaks in spectra derived from MALDI mass spectrometric measurements were annotated as follows: In order to determine the instrument baseline, the mass spectrometric data was two times filtered using a low-pass median parametric spline filter. The smoothed residual mean standard deviation from the baseline is used as an

estimate of the instrument noise level in the data. After baseline correction and rescaling of the data in level-over-noise coordinates, the data point with the largest deviation from the baseline is used to seed a non-linear (Levenberg-Marquardt) data fitting procedure[32] to detect possible peptide peaks. Levenberg-Marquardt is an alternative to the Gauss-Newton method of finding the minimum of a function *F(x)* that is a sum of squares of non-linear functions. Specifically, the fit procedure attempts to produce the best fitting average theoretical peptide isotope distribution parameterized by peak height, resolution, and monoisotopic mass. The convergence to a significant fit is determined tracking sigma values[32]. Convergence is reached if sigma does not change more than 0.1 for five successive iterations. After a successful convergence, an estimate for the errors of the determined parameters is produced. This is done by applying a bootstrap procedure using sixteen repeats, for each repeat exchanging randomly 1/3 of the data points. The resulting fit is subtracted from the data, the noise level in the vicinity of the fit is adjusted to the sum of the extrapolated noise level and the deviation from the peak fit. The process is iterated to find the next peak as long as a candidate peak more than five times over level of noise can be found. The peak annotation is stopped when 50 data peaks have been found. The zero and first order of the time-of flight to mass conversion are corrected using linear extrapolation from detected internal standard peaks, and confidence intervals for the monoisotopic mass values are estimated form the mass accuracies of the peaks and standards.

3.3. Probabilistic matching of spectra peaks

Peak mass lists for mass spectra are directly compared to theoretical digests for whole protein sequence databases. For each theoretical digest, $\left[1-\prod(1-NP(p_i))\right]^{cMatches}$ is calculated[33]. In this formulation N is the number of peptides in the theoretical digest, $P(p_i)$ is the number of peptides that match the confidence interval for the monoisotopic mass of the peak divided by the count of all peptides in the sequence database, and *cMatches* is the number of matches between digest and mass spectrum. It can be shown that this value is proportional to the probability of obtaining a false positive match

between digest and spectrum. Probability values are further filtered for high significances of the spectra peaks that produce the matches. After a first round of identifications, deviations of the identifications for mass spectra acquired under identical conditions are used to correct the second and third order terms of the time-of-flight to mass conversion. The resulting mass values have mostly absolute deviations less than 10 ppm. These mass values are then used for a final round of matching, where all matches having a P_{mism} less than *0.01/NProteins* are accepted.

4. Algorithms and Methods

4.1. Spectra similarity

4.1.1. Derivation of the spectra similarity

All spectra we analyzed have been treated as described in chapter 3. The peaks have been annotated and the spectra with its peaks have been calibrated using internal standards. After completion of these procedures, we assumed that all deviations from true values in the data were due to non-systematic deviations. Thus we can treat the measured monoisotopic mass value as a sampling from a normally distributed population of possible mass values for a measurement.

The probability density of the distribution of monoisotopic masses is therefore the Gaussian function.

A Gaussian[34] is defined as follows: $f(x) = ae^{\frac{-(x-m_0)}{2s^2}}$ (1)

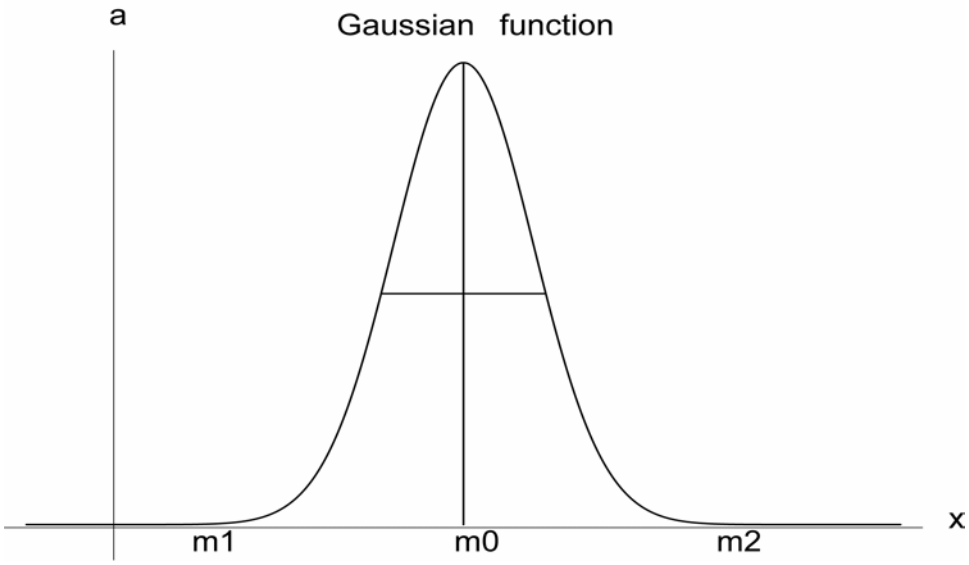


Figure 2: This Figure shows a Gaussian shape of a peak with monoisotopic mass m_0. Full width at half maximum is defined as $\pm s\sqrt{2\ln(2)}$. The amplitude of the curve is defined by a.

s denotes the standard deviation, which is defined as follows:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{i=N}(m_i - m_0)^2} \quad (2)$$

The discrete correlation theorem states that the discrete correlation of two real functions g and h is one member of the discrete Fourier transform pair Corr(g,h) ⇔ $G_k H_k^*$ where G_k and H_k are discrete Fourier transforms of g_j and h_j, and the asterix denotes complex conjugation[35]. Therefore, the correlation between two spectra is the inverse Fourier transform of the product of the Fourier transform of the first spectrum with the complex conjugate of the Fourier transform of the second spectrum. Foruier transforms in signal processing are usually evaluated using the well known Fast Fourier Transform (FFT) algorithm. However, calculating FFT's with resolutions of <10 ppm requires millions of data bins and is not very efficient. Since it is obvious that a Gaussian transformed in Fourier space again yields a Gaussian, we decided to calculate the correlation explicitly (see Appendix). This can be used to yield the mass correlation function *mc(x)* for sums of Gaussians. The following formulation denotes the mass correlation function *mc(x)*.

$$f(x) = \sum_{i=1}^{i=N_i}\sum_{j=1}^{j=N_j} a_i a_j s_i s_j \frac{1}{\sqrt{s_i^2 + s_j^2}} e^{\frac{-(m_{0i} - m_{0j} - x)^2}{2(s_i^2 + s_j^2)}} \quad (3)$$

Formula 3 is the final formulation of the mass correlation function that treats the amplitudes of the peaks as height *a* and is depending on the size of the standard deviations s_i and s_j respectively as well as the difference in monoisotopic masses m_{0i} and m_{0j} respectively. As it is observed, the exponent is getting large as the difference in mass of the two compared mass grows. Thus, large differences in the compared monoisotopic masses result in very small correlation values. The exponential term reaches its maximum of 1 when the difference in monoisotopic masses is equal to zero, and so does the function in general, when the two peaks perfectly match.

In this formulation, each peak (*i*) from one spectrum is compared to all the peaks (*j*) of the other spectrum. Spectrum one contains *Ni* peaks and spectrum two *Nj*. The exponent becomes zero when the monoisotopic masses (m_0) of peak *i* and *j* are exactly the same at a lag $x = 0$. s_i and s_j denote the standard deviations of the peaks. For

calibrated mass spectra, as it is always the case regarding the data we analyzed, the lag x is zero per definition. However, we derived the formula for the general case. The above formalism also includes the amplitudes of the compared peaks represented by a_i and a_j respectively. We can treat spectra assuming that all the peaks have the same intensity. Thus, the height a_i and a_j respectively are set to 1. Formulation (3) then yields

$$mc(x) = \sum_{i=1}^{i=N_i} \sum_{j=1}^{j=N_j} s_i s_j \frac{1}{\sqrt{s_i^2 + s_j^2}} e^{\frac{-(m_{0i} - m_{0j} - x)^2}{2(s_i^2 + s_j^2)}} \quad (4),$$

which denotes the actual mass correlation function *mc(x)*.

4.1.2. Normalizing the mass correlation function

As it is realized when calculating the autocorrelation, the mass correlation in its raw form is not normalized. The auto-correlated form yields a sigma dimension, which means that the correlation function has to be made dimension less in order to yield similarity values normalized from 0 to 1. The correlation function is usually normalized by dividing it with the geometric average of the two individual autocorrelations[36]. This formulation is shown on equation (1).

$$corr(i,j,x) = \frac{\sum_{i=1}^{i=N_i} \sum_{j=1}^{j=N_j} s_i s_j \frac{1}{\sqrt{s_i^2 + s_j^2}} e^{\frac{-(m_{0i} - m_{0j} - x)^2}{2(s_i^2 + s_j^2)}}}{\sqrt{\sum_{i=1}^{i=N_i} \sum_{i=1}^{j=N_j} s_i s_i \frac{1}{\sqrt{s_i^2 + s_i^2}} e^{\frac{-(m_{0i} - m_{0i} - x)^2}{2(s_i^2 + s_i^2)}}} \sqrt{\sum_{j=1}^{j=N_j} \sum_{j=1}^{j=N_j} s_j s_j \frac{1}{\sqrt{s_j^2 + s_j^2}} e^{\frac{-(m_{0j} - m_{0j} - x)^2}{2(s_j^2 + s_j^2)}}}} \quad (1)$$

To calculate the normalization factor, only overlapping peaks are taken into account. The normalizing factor yields then:

$$\frac{1}{\sqrt{\sum_{i=1}^{i=N_i} \sum_{i=1}^{i=N_i} \frac{s_i}{\sqrt{2}}} \sqrt{\sum_{j=1}^{j=N_j} \sum_{j=1}^{j=N_j} \frac{s_j}{\sqrt{2}}}} = \frac{\sqrt{2}}{\sqrt{\sum_{i=1}^{i=N_i} \sum_{i=1}^{i=N_i} s_i} \sum_{j=1}^{j=N_j} \sum_{j=1}^{j=N_j} s_j} \quad (2)$$

The normalization factor contains an inverted sigma term that is multiplied by the square root of 2. The normalized mass correlation is then the product of the mass correlation with the normalization factor. This formulation looks as follows:

$$mc(x) = \frac{\sqrt{2}}{\sqrt{\sum_{i=1}^{i=N_i}\sum_{i=1}^{i=N_i} s_i \sum_{j=1}^{j=N_j}\sum_{j=1}^{j=N_j} s_j}} \sum_{i=1}^{i=N_i}\sum_{j=1}^{j=N_j} s_i s_j \frac{1}{\sqrt{s_i^2 + s_j^2}} e^{\frac{-(m_{0i} - m_{0j} - x)^2}{2(s_i^2 + s_j^2)}} \quad (3)$$

By setting $i = j$, which evaluates the autocorrelation, it can be seen that now the normalized form lost its sigma dimension and results to one. If none of the peaks from either spectrum overlap when calculating the similarity between two different spectra, the mass correlation yields zero. Thus, the minimum value for similarity is 0 and the maximum value is 1. The formalism shown in formula (3) can be further simplified. As we are only interested in the fraction of peaks from two different spectra that are overlapping and assuming that the spectra are calibrated (lag $x = 0$), the formula can be reduced to:

$$mc(x) \cong \frac{\sqrt{2}}{\sqrt{\sum_{i=1}^{i=N_i} s_i \sum_{j=1}^{j=N_j} s_j}} \sum_{i,j=1}^{i,j=Noverlap} s_i s_j \frac{1}{\sqrt{s_i^2 + s_j^2}} e^{\frac{-(m_{0i} - m_{0j})^2}{2(s_i^2 + s_j^2)}} \quad (4)$$

$N_{overlap}$ is the fraction of peaks that overlap. Overlapping peaks are two peaks from two different spectra that are defined by their monoisotopic masses m_{0i} and m_{0j} respectively that are within two times of their confidence interval $\left|m_{0i} - m_{0j}\right| \le 2\left(\frac{s_i + s_j}{2}\right)$.

An estimate of the error of this simplification is given as follows: Consider a case where two spectra are compared each having 50 peaks. The peaks of one spectrum are all shifted by 1.0 Da and the standard deviation of the peaks is 0.2 for all peaks. The resulting correlation value for these two spectra is ~0.015. This is a very weak correlation value even though the masses are shifted by only 1.0 Da. Given the fact that the distribution of peaks in real spectra is much broader, the error of the simplified mass correlation function is small and thus, can be neglected.

We have already stated that the individual peak intensities have been consciously excluded from the correlation function. As it is obvious from Formula 4 of the previous chapter, the contribution of the peak intensities amounts to the cross product of amplitudes for overlapping peaks (and can be normalized in the usual way to unit vector in the N_{peak}-dimensional space). The contribution of amplitudes is

multiplicatively connected to the correlation of masses. It is shown in the results section that the inclusion of the peak heights decreases the robustness of the correlation function. In the case of MALDI mass spectrometric data, reproducing exact intensities over a whole spectrum is very difficult, as slight changes in laser power, acquisition parameters, or crystallization parameters can alter the spectrum. Instead of neglecting the intensity of a peak, we replaced the information about the height of a peak by its ordered rank. We therefore introduced the more robust rank order correlation coefficient, described by Spearman [37].

4.1.3. Spearman rank order correlation coefficient

The concept of rank correlation is the following: We are given N pairs of measurements (x_i and y_i). If we replace the value of x_i by the value of its rank among all the other x_i's in the sample, that is, *1,2,3,...N*, then the resulting list of numbers will be drawn from a perfectly known distribution function, namely uniformly from the integers between 1 and N, inclusive [37]. If some of the ranks have identical values, it is conventional to assign to all these "ties" the mean of the ranks that they would have had if their values had been slightly different. These tied ranks are called midranks. The sum of all assigned ranks will be the same as the sum of the integers from 1 to N, namely $\frac{1}{2}N(N+1)$. The same procedure is done for the y_i's, replacing each value by its rank.

A statistical measure of the rank correlation has been described by Spearman. The Spearman rank order correlation is defined by

$$r_c = \frac{\sum_i (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_i (R_i - \overline{R})^2}\sqrt{\sum_i (S_i - \overline{S})^2}} \quad (1)$$

R_i is the rank of peak i in spectrum one and S_i is the rank of peak i in spectrum two. $\overline{R}$ and $\overline{S}$ are the midranks. In our procedure, if two peaks have intensities that are within 10% of each other, the same rank (tie) is assigned and for all ties, the corresponding midrank is calculated. The exact relation of the ranks including ranks with the same value is shown in the following formula (2).

$$r_s = \frac{1 - \frac{6}{N^3 - N}\left[D + \frac{1}{12}\sum_k (f_k^3 - f_k) + \frac{1}{12}\sum_m (g_m^3 - g_m) \right]}{\left[1 - \frac{\sum_k (f_k^3 - f_k)}{N^3 - N} \right]\left[1 - \frac{\sum_m (g_m^3 - g_m)}{N^3 - N} \right]} \quad (2)$$

f_k is the number of ties in the kth group of ties among the R_i's, and g_m is the number of ties in the mth group of ties among the S_i's. If all the f_k's and all the g_m's are equal to one, meaning that there are no ties, then equation (2) reduces to equation (3).

$$r_s = 1 - \frac{6D}{N^3 - N} \quad (3)$$

D , the so called sum-squared difference in ranks, is closely related to r_s and is defined as

$$D = \sum_{i=1}^{N} (R_i - S_i)^2 \quad (4)$$

4.1.4. Weighting of the correlations

Mass and rank correlation are supposed to be orthogonal (as they use different information from the spectrum). Thus, they should be equally weighted. We have evaluated the exact relation between the two correlations by the following formulation,

$$cv = e^{\frac{k\log(rc) + \log(mc)}{1+k}} \quad (1),$$

where k is the weighting factor, rc the rank correlation and mc the mass correlation. By assigning the weighting factor k values from 0 to 2, we evaluated different weightings of the two correlations. The most accurate relation is shown to be at $k = 1$, which means the two correlations have to be equally weighted (see results section). Setting k to 1 yields the correlation value to become the geometric mean of the two correlations.

$$cv = \sqrt{mc \cdot rc} \quad (2)$$

4.2. Hierarchical clustering methods

We used our newly established similarity measure to compare large datasets of experimental mass spectrometric data. All the spectra compared have been measured and treated as described in section 3. We have been employing two different kind of hierarchical clustering methods, trees and graphs, to gather spectra together. Spectra are clustered together based on the score derived from pair wise comparison of the spectra using the correlation function described in section 4.1.

4.2.1. Tree-based clustering

Among the many possible methods to generate a tree from pairwise similarity scores, we chose a method that builds a tree by iteratively calculating the average distance between nodes and leafs. Other tree building algorithms differ only in the way of calculating the distance between two clusters. We used the method that is described in detail below as a proof of concept.

This clustering procedure described by *Sokal and Michener* [38] is called UPGMA, which stands for unweighted pair group method using arithmetic averages. It works by clustering the spectra, at each level amalgamating two clusters and at the same time creating a new node on the tree. The tree can be imaged as being assembled upwards and each node being added above the others, and the edge lengths being determined by the difference of the heights at the bottom and the top of an edge[39].

The distance d between two clusters C_i and C_j is defined as the average distance between pairs of spectra from each cluster.

$$d_{ij} = \frac{1}{|C_i||C_j|} \sum_{p \to C_i, q \to C_j} d_{pq} \quad (1)$$

$|C_i|$ and $|C_j|$ denote the number of spectra in clusters i and j, respectively. If C_k is the union of the two clusters C_i and C_j ($C_k = C_i \cup C_j$) and C_l is any other cluster, then the distance between the two clusters is defined by:

$$d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_i| + |C_j|} \quad (2)$$

The clustering procedure is:

Initialization:

Assign each spectrum *i* to a cluster C_i.

For each spectrum define one leaf of the tree and place it at height zero.

Iteration:

Determine the two clusters *i, j*, for which the distance d_{ij} is minimal.

Define a new cluster *k* by $C_k = C_i \cup C_j$, and define d_{kl} for all *l* according to formula (2).

Define a node *k* with daughter nodes *i* and *j*, and place it at height $d_{ij}/2$

Add *k* to the current clusters and remove *i* and *j*

Termination:

When only two clusters *i, j* remain, place the root at height $d_{ij}/2$.

A detailed cartoon of the UPGMA clustering algorithm and tree construction is given on Figure 2. Once a complete tree of the sample set is built, we can begin to disjoint it back into pieces, equivalent to generate sub trees or clusters from the whole tree structure. We used a threshold parameter in order to cut a tree separate clusters. This process is evaluated iteratively by checking the cluster consistency using a user supplied estimate of cluster count and the total number of spectra clustered.

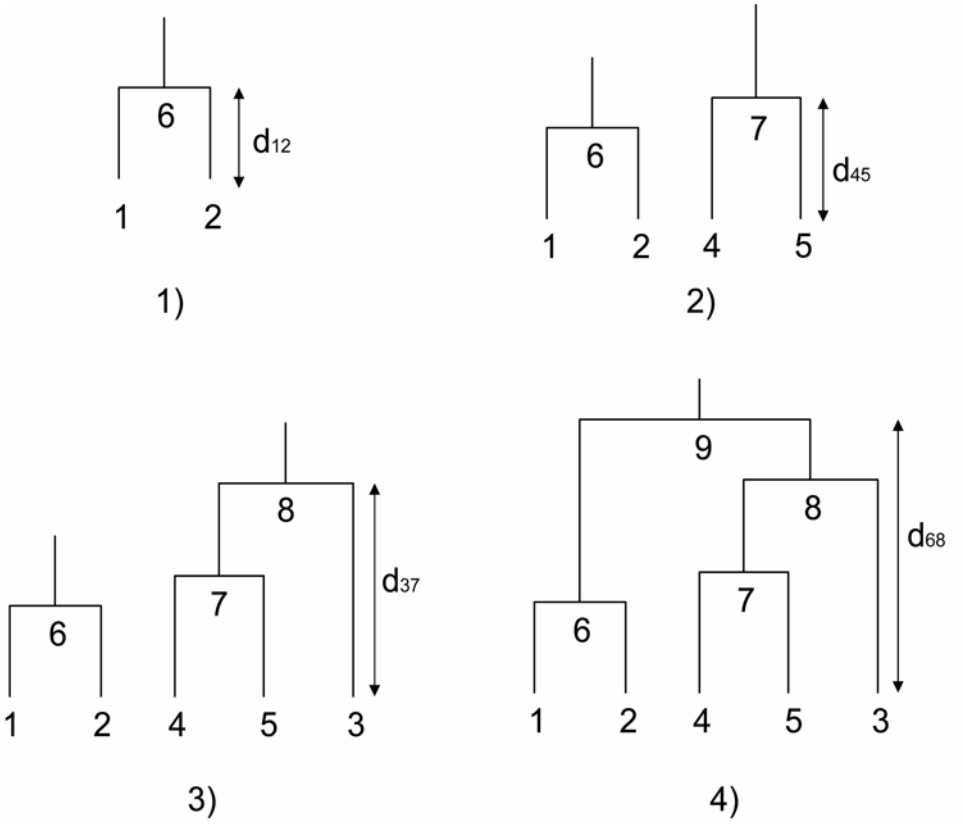


Figure 3: This figure demonstrates how UPGMA produces a rooted tree by clustering spectra. An example set consisting of five spectra is given. Distances between spectra are measured using the spectra similarity algorithm described in chapter 4.1, while the UPGMA algorithm derives distances between clusters.

4.2.2. Graph theoretical clustering

A graph is defined as a set of nodes and a set of lines that connect the nodes. We call V the *vertex set* and E the *edge set* of G. The degree of a node is the number of nodes that a given node is connected to[40].

Using the correlation values, we construct a graph by connecting spectra with the highest available correlation value from a set of unconnected nodes that will not connect to a spectrum that has already been linked to spectra in the current component of the graph. As our means of constructing a graph includes only highly similar spectra not yet clustered, the resulting graph is a non-circular graph.

Thus, our procedure constructs a forest of minimal spanning trees with primitive natural chain breaks. A minimal spanning tree data structure allows walking along the graph, from the shortest distance between two nodes to the farthest node, visiting each

node at least once. The algorithm we employ for the clustering of spectra sets also reduces the dimensionality of the problem from two dimensions (the trigonal matrix) to one dimension, allowing for highly efficient processing of large amounts of data[41].

The clustering procedure is:

Initialization:

Construct of the trigonal matrix $n*(m-1)/2$

Evaluate the similarity measure for each pair of spectra where $n < m$.

Construct of a list of connections according to the following procedure: For column i in the trigonal matrix pick $i \rightarrow k$ with *max (R)* and add it to list *L(i,j)*.

Iteration:

Build the graph by iterating through list *L* and connect the nodes stepwise.

Edges are only taken into account when the similarity value is larger than a specified threshold.

Termination:

The graph(s) are constructed when all the nodes contained in list L are connected.

We argue that this list describes one or more directed graphs with the following properties:

1) It contains only nodes that have at least one connection.

2) Each outgoing connection points to a higher target node number.

3) It does not contain paths that are circular.

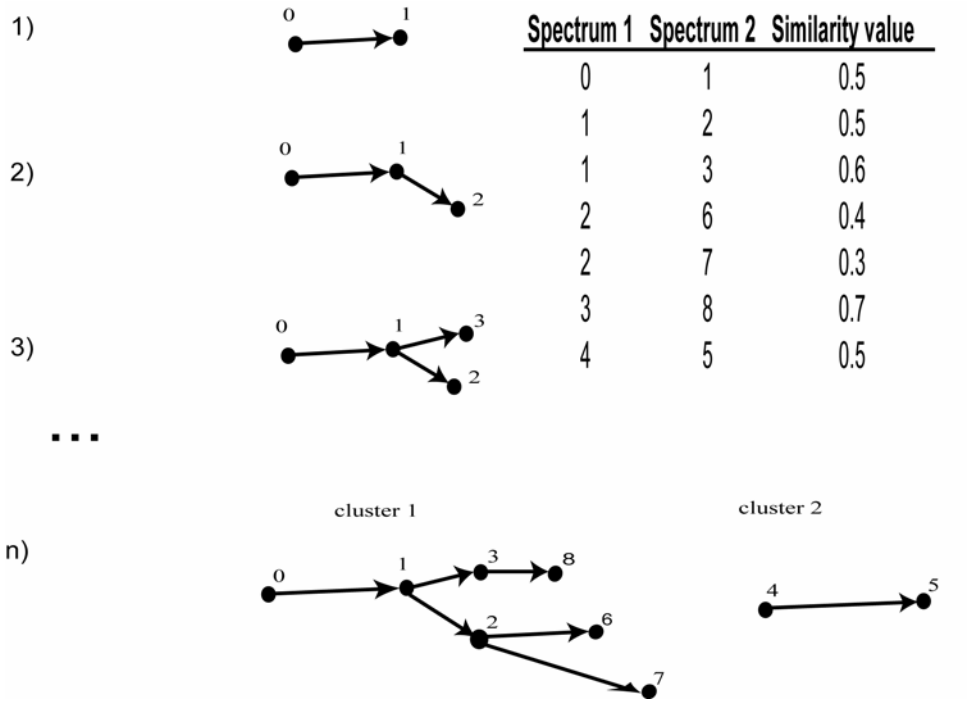
The non-circularity is shown as follows: If we assume that it would be possible to construct a circular graph under these conditions we could find the node with the smallest node number that has an incoming connection. Assuming circularity, we could reach this node from another directly connected node. This node has to have a node number greater than the node number of the original node. However, for this node the connection to the original node would be an outgoing connection, and would have to point to a node with a higher node number. Therefore, construction of circular paths is impossible if conditions 1 and 2 are to be met.

As only the highest scorers (spectra that have no spectrum that is more similar to any of them in the spectra set) get the chance to be connected, we will obtain clusters that are maximal consistent. Nevertheless, we encountered problems in the grouping process due to spectra that contained peptide peaks from two different proteins. Protein mixture spectra act as links between two independent clusters. In order to prevent such inconsistent cluster formation due to protein mixtures we have introduced an optional variable threshold parameter. This parameter setting is varied in an iterative process where the cluster consistency is evaluated each time. Iteration is stopped when all resulting clusters are uniform.

A simple example of the graph constructing algorithm is shown on Figure 4. Table 1 gives an example of a trigonal matrix calculated from a sample set of 10 spectra. The correlation values are obtained by evaluation of the similarity algorithm. The trigonal matrix is transformed into a list of most similar pairs of spectra. This list is then evaluated in order to build the graph.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	0.5	0.4	0.4	0	0	0.3	0.2	0.4	0	0
1		0.5	0.6	0	0	0.3	0.2	0.5	0	0
2			0.5	0	0	0.4	0.2	0.5	0	0
3				0	0	0.3	0.2	0.6	0	0
4					0.5	0	0	0	0	0
5						0	0	0	0	0
6							0.2	0.4	0	0
7								0.2	0	0
8									0	0
9										0

Table 1: This Figure shows a trigonal matrix generated from pair-wise comparisons of 10 spectra. The highest correlation value per column is marked and added to list *L(i,j)*.



Spectrum 1	Spectrum 2	Similarity value
0	1	0.5
1	2	0.5
1	3	0.6
2	6	0.4
2	7	0.3
3	8	0.7
4	5	0.5

Figure 4: The graph is build by iteration through list *L(i,j)* and successively addition of new nodes to the graph. Only vertices with a value of 0.3 and above are drawn.

4.3. Generation of consensus spectra

The clustering process resulted in a list of clusters each representing a group of similar spectra from the sample set. The ideal case would be one cluster for every protein. The latter is frequently not realized because there are spectra that contain peptides from more than one protein. These protein mixture spectra denote link clusters containing spectra describing each part of the mixture spectrum. Clusters linked via mixture spectra are usually large and contain both, mixture spectra and "pure" spectra.

An obvious exertion of a database of clusters would be the generation of consensus spectra which contain the most abundant peaks occurring in measured spectra as assembled in each cluster. Each resulting consensus spectrum is the representative spectrum for this kind of peptide assembly. For example, a cluster that contains spectra that are eventually identified as protein X would yield a consensus spectrum that is representing a robust and averaged empirical mass spectrometric representation of protein X.

We experimented with different definitions of consensus spectra and came to the decision to take the 50 most abundant overlapping peaks that occur in spectra from the cluster. As we are able to collect more than one experimentally measured monoisotopic mass per peptide, the average of the experimental monoisotopic masses is more exact than each individual sampling. Additionally, we are able to define realistic standard deviations for the peak masses, as the standard deviation can now be estimated from statistical sampling, as opposed to the assumption that the measured variance in peak definition is the main determinant of deviations for monoisotopic masses.

More accurate standard deviations play an important role in peptide mass fingerprinting protein identification – they allow to sample a realistic set of peptides from a mass range. However, probably the biggest advantage is that we are able to define a rank order that reflects the real rank of any specified peptide as accurate and as robust as possible.

Algorithm for the construction of consensus spectra:

Initialization:

Pair–wise comparison of all spectra in a cluster

Create a list $L(p_i)$ of overlapping peaks. An overlap occurs when $\left|m_{0i} - m_{0j}\right| \leq 2\left(\frac{s_i + s_j}{2}\right)$. This will miss less than 1% of the real overlaps.

Calculate average masses, new standard deviations, and occurrence of these peaks.

Iteration:

Iterate through list $L(pi)$ and compare each mass m_{ij} to all the other masses in the list.

Iterate through all the spectra and score the order of each pair of masses according to their height relation in the spectrum. If the height of peak with mass i is larger than the height of peak with mass j, it is scored 1, if it is smaller it is scored -1. If this pair does not occur or if they are tied in height the relation is scored 0.

Termination:

Calculate the sum and sort descending according to the sum. The highest score is ranked first, second highest ranked second and so on.

4.4. Database of consensus spectra

We generated datasets of totally 100,000 spectra and evaluated them as described in the previous chapters. The samples were derived from four different organisms:

~ 35000 spectra from *Human HEK293* cell line

~ 15000 spectra from *B. subtilis*

~ 14000 spectra from *Paracoccus z.*

~ 18000 spectra from *Rat* insulinoma cell line *INS1*

~ 18000 spectra from *Human blood plasma*

The cell line samples used came from conventional, untreated, log-phase cultures. All spectra have been treated as described in chapter 3.

We clustered MALDI mass spectra from different sample sets and generated consensus spectra out of the resulting clusters as described in section 4.3. All consensus spectra that were derived from clusters containing more than two spectra were analyzed by peptide mass fingerprinting and peak peptide matching after identification. Peak peptide matching is a procedure which tries to assign additional candidate sequences to peaks in spectra that are identified. These peptides include miscleaved peptides, modified peptides, as well as peaks that have standard deviations too large to be matched. All the identified consensus spectra were stored in a database along with the matched and the unmatched peptides. A matched peptide is a peptide that has been found to match the monoisotopic mass of a peak in peptide mass fingerprinting or in peak/peptide matching. On the other hand, unmatched peptides denote peptides that could not be assigned to peaks in the consensus spectrum by employing either method. In general, we see only forty percent of the theoretically occurring peptides in a spectrum.

In Chapter 2 we described the need for a realistic peptide set for peptide mass fingerprinting in order to reduce the number of possible peptide matches. It was mentioned that in terms of probability it is favorable to consider only a minimal set of peptides. As stated above, only 40% of the peptides in theoretical digests are observed. Therefore, there should be a reasonable potential for rules that can recognize peptides that should not occur in practice.

4.5. Peptide Analysis

We used the previously described database of consensus spectra to compare the matching peptides with the unmatched in order to test different peptide properties like missed cleavages, hydrophobicity, hydrophobicity gradient, and distributions of amino acid in peptide sequences.

4.5.1. Missed cleavages

For the analysis of missed cleavages, we used consensus spectra from clustering of *Bacillus subtilis* and *Human* samples. With the peak lists of the consensus spectra we carried out database searches in order to identify the underlying proteins. This first identification cycle was done without considering missed cleavages or posttranslational modifications. If a spectrum was identified, it was compared to the theoretical digest of the identified protein again, allowing this time two missed cleavages within the sequences. The resulting peak / peptide matches were recorded and analyzed for predictive features.

4.5.2. Hydrophobicity and hydrophobicity gradient

The hydrophobicity of a peptide was calculated by summing per amino acid hydrophobicity scores according to *Kyte and Doolittle*[42]. This procedure scores each amino acid from a given sequence with scores ranging from 0 for Arginine to 9.0 for Isoleucine. We also evaluated the hydrophobicity gradient of the peptide sequences by estimating the slope of the hydrophobicity score along the sequence.

4.5.3. Amino acid distribution in peptide sequences

To obtain information on the amino acid distribution at every position in a peptide sequence, we calculated the relative entropy of each amino acid at every position. Relative entropy can be seen as information content in ‘bits’ when the distribution of

the experimentally determined amino acid frequency at a certain position is compared to its background distribution e.g. the amino acid composition of the SWISS-PROT database[43]. The concept of entropy was described by Shannon[44]. It can be used to measure the degree of conservation at a site in a peptide sequence alignment. For two distributions *P* and *Q* the relative entropy is defined by

$$H(P \parallel Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)} \quad (1),$$

where $P(x_i)$ is the measured amino acid frequency and $Q(x_i)$ describes the background frequency. The entropy of a given distribution is always positive.

4.6. Noise reduction in mass spectrometric data

Despite the quality of mass spectrometric measurements using today's technical equipment, there are still components in the spectra that are not derived from measured proteins. These components are called noise components. Unfortunately, they interfere with identifications carried out by peptide mass fingerprint algorithms. Noise components are known peaks derived from enzymatic cleavage (autolysis peaks), matrix peaks and other substances, including polymers or peaks derived from improper treatment of measured samples (e.g. kreatine peaks). All peaks that disturb the spectrum should be excluded in order to decrease the probability false positive assignments.

We describe two different approaches to clean the spectra from noise. By applying the first approach, spectra are filtered by information extracted from the spectrum itself. That means that no additional information is needed. The second approach uses additional information given by images of 2D-gels and spot locations mapped on them. It is clear that the second method is applicable only for experiments that used the separation of proteins by 2D-PAGE.

4.6.1. Noise filtering on a single spectrum

While it is trivial to eliminate noise peaks that regularly appear in spectra (for example the mentioned kreatine peaks whose exact monoisotopic mass is known), it is slightly more complicated to filter noise peaks with unpredicted monoisotopic masses (e.g. matrix peaks or polymer peaks).

To detect polymer peaks in the spectrum it is possible to make use of the oligomeric nature of polymers. Peaks from polymer chains of various lengths are separated by regular, repeating mass differences (usually the mass of the monomer). The technique of choice to detect adjacent peaks with a defined difference in mass is to use autocorrelation (correlation of a spectrum with itself). To search for polymer specific differences, the autocorrelation has to be calculated for a range of different shifts in masses. For example, to detect a polymer with a monomer mass of ~22, the spectrum for which the autocorrelation is to be evaluated has to be shifted by a lag of 22 to detect a peak in the resulting correlation evaluation. In order to detect potential polymers, we evaluated the autocorrelation for a spectrum about a hundred times, each time increasing the mass shift of the spectrum by one. The frequency of polymer peaks (the mass of a monomer) is than observed by plotting the autocorrelation versus the mass shifts. Once the mass-difference is known, polymer peaks in the spectrum can be found by searching for chains of this mass difference between peaks.

4.6.2. Noise filtering using information from 2D-PAGE

Matrix peaks and small chemical noise peaks are by far the most abundant species that should be filtered out of the spectrum. The monoisotopic masses of these noise peaks are often not known and therefore cannot be eliminated a priori. As it is impossible to distinct such noise peaks in a single spectrum, we were looking to use statistical properties of a set of several samples to find out suspect peaks. Reproducible assemblies of noise peaks will cluster into one or more consensus spectra per definition. While it is impossible to differentiate these consensus spectra based on their mass information alone, it is well possible to compare the spatial distribution of the spots the

spectra were derived from with the spatial distribution of all spots excised from the gel. Real protein spots should show a spatial distribution that is inconsistent with random sampling from the spot set, while noise spectra should be distributed as a random sampling from the spots. To differentiate between noise clusters and "good" clusters, we compare the distribution of spots the spectra were originated from to the distribution of the spectra in the clusters. While the randomness of the sampling could be shown using exact statistical criteria (Komolgorov-Smirnov)[45], by a simple calculation of the area where the spots originally have been excised, we are able to determine whether the cluster is a noise cluster or not.

This is illustrated schematically on Figure 5. The resulting consensus spectrum of such a cluster contains a majority of peaks that are noise peaks. By excluding these peaks from the dataset, the spectra in general become cleaner and the overall clusterability of the spectra is improved.

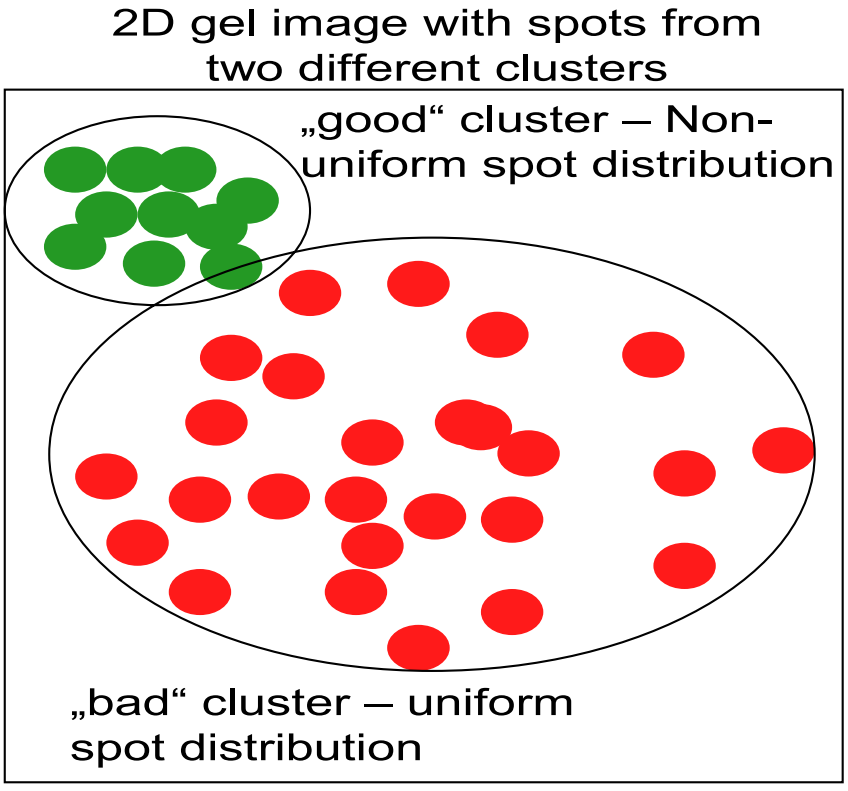
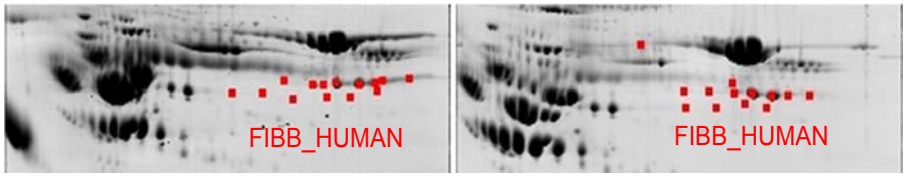


Figure 5: This figure illustrates how we are able to tell whether a cluster is a noise cluster or not by comparison of the spot distributions. As the spots from the bad cluster are uniformly distributed, the region cannot be deterministic for a single protein. Peaks in consensus spectra from noise clusters can be excluded from the dataset.

4.7. Use of clusters of mass spectra for gel image analysis

We compared the assemblies of spots derived from their peptide mass fingerprinting identifications with the information on the spot locations derived from the clustering of similar spectra. As already stated, we know the location of every spot on the gel and therefore, we know for every spectrum where it has been measured from. Such an analysis is shown on Figure 6, illustrating all spots for which the measured spectra have been identified as *fibrinogen beta chain precursor*. This Figure also shows the spots whose spectra were clustered together using the clustering algorithm and similarity measure described in chapter 4.1. The resulting consensus spectrum of this cluster was also identified as *fibrinogen beta chain precursor*. As shown on the Figure, there is a strong correspondence between the results of peptide mass fingerprint identification and clustering, however, the clustering procedure identifies a more comprehensive set of spots.

Spectra from these spots have been identified as *fibrinogen beta chain precursor* by peptide mass fingerprinting.



Spectra from these spots have been clustered together and the resulting consensus spectrum has been identified as *fibrinogen beta chain precursor*.

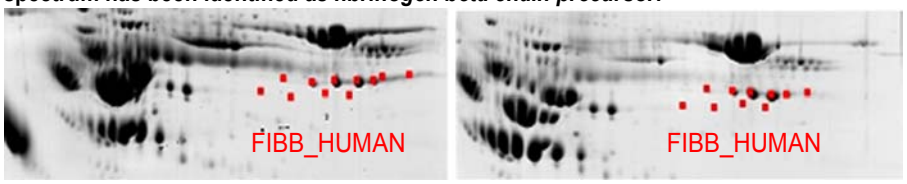


Figure 6: This figure illustrates the mapping of spectra to their spots resulting from clustering and from peptide mass fingerprint identifications. The lower two gels show the spots of spectra assembled in a single cluster whose consensus spectrum has been identified as *fibrinogen beta chain precursor*. The upper gels show spots whose spectra have been identified as *fibrinogen beta chain precursor*. On the upper gel images, some outlying spots, probably protein mixture spectra were not clustered with the "pure" *fibrinogen beta chain precursor* spectra.

Mapping the cluster information onto the gel(s) can be done for all the clusters. To aid visualization of gel regions identified by the clusters, we used the Voronoi[46] algorithm to partitionate the gel and its spots into separate areas. This algorithm spans a network

of polygons on the gel, with each spot having its own area that is not intersected by another area.

As spectra information is rarely complete, not every spot on the gel will be included into an identified region. Missing spot information is then added in two ways. First, polygons or its spots are added to a cluster defined region when the undefined polygon is fully surrounded by polygons that are assigned to the same cluster. We assume in that case that the missing spot belongs to the same cluster.

While this is certainly possible, only few spots can be assigned using this simple approach.

A more powerful way of complementing missing spot information is to draw on redundancies between replicate gels. With the second method, we try to add missing spot information by searching and comparing the immediate surrounding of a non-clustered spot to surroundings of clustered spots on other gels. We employed a simple k-Means[47] method to determine whether a non-clustered spot can be assigned to a cluster, of which similar clusters surround the same region. The method is described as follows: After creation of a raw picture with all the clusters on the gel, we generate a list of surrounding polygons of all non-clustered spots. In order to add the non-clustered spots to a cluster, the surroundings of all the other spots located on different gels are searched for the best matching neighborhood. If there is such a match, the non-clustered spot is assumed to belong to the same cluster as the nearest clustered neighbor of that matching surrounding. A sketch of this method is given on Figure 7.

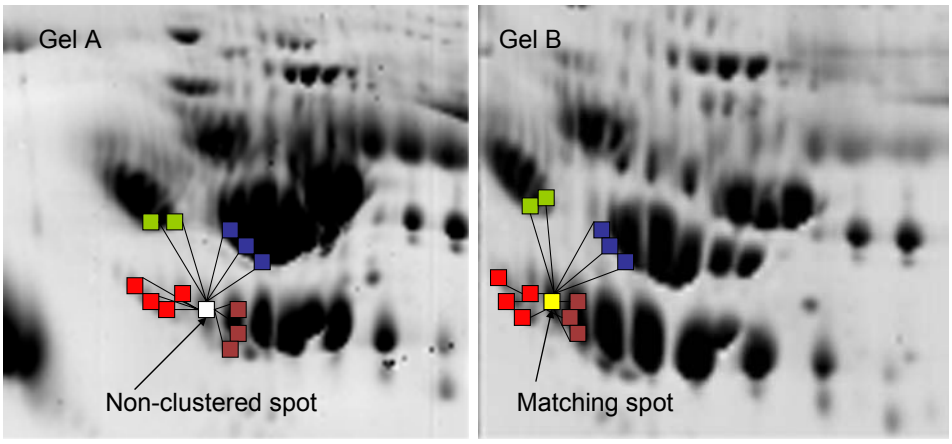


Figure 7: This figure demonstrates the use of k-means criterion to decide whether two surroundings match or not. Same color means same cluster and therefore same protein.

The neighborhood found to match the surrounding of the non-clustered spot on gel B is used to identify the spot on gel A.

4.8. Comparative gel set analysis

The comparative approach outlined above can be extended to difference analysis of sets of gels, a very frequently employed approach to proteomics data processing.

In the following chapter, it is explained how we are able to compare two different sets of spectra (so called gel/spectra sets) accurately, and how we can extract valuable information in terms of differences in expressed proteins or even changes in protein expression levels.

A gel set is defined as collection of 2D-gels showing protein expression of a single sample. These sets include different slices of the proteome under investigation as derived by sample preparation, which are measured usually in replicates. All the spots from gels in a set will be automatically excised and tryptically digested, and resulting peptide mass fingerprints measured two or more times on MALDI-TOF mass spectrometers. It is important that gel sets contain many redundant gels. This redundancy is the statistical background that allows such an analysis.

Consider the simple case where two gel sets were generated from the same tissue, with the difference that one sample has been treated with a substance (e.g. a drug), whereas the other sample acts as a control. Traditionally employed methods to compare such gel sets were based on the analysis of differences in protein expression, as observed by gel image comparisons.

With our method to compare the two gel sets, their spectra have to be clustered separately and a consensus spectrum has to be generated for every resulting cluster. In a second step, all the regions on the gels, assigned to the same consensus, have to be calculated to evaluate the sum of spot volumes for every region. In order to find similar regions on gels that belong to another gel set, all the consensus spectra from one gel set are compared to the consensus spectra of the other gel set. The result of this comparison consists of three sets of consensus spectra. The first set contains consensus spectra that have at least one similar consensus spectrum in the compared gel set(s). For these matching spectra, the degree of change in spot volumes can be calculated from spot

intensities on the gel images, which gives an estimate for protein expression differences. The second and third set contains consensus spectra that are unique in the sense that there is no similar counterpart found in either of the gel sets. This could be due to insufficient coverage by PMF measurements, or due to large differences in expression levels. However, sampling of more than three replicates will reduce the chances of annotating a false positive expression difference to less than 5% if the identification rate is larger than 70%. Only after the pair–wise comparison of the consensus spectra, spectra are identified using the peptide mass fingerprinting technique or using the reference spectra library.

Therefore, we established a method to compare large datasets of spectra that is independent from identifications of the spectra and independent from gel image comparisons. We measure the differences directly on the level of mass spectrometric information, which results in increased accuracy of comparisons and makes the conclusions of the analysis more robust. By addition of the spot intensities, we are able to receive a raw estimate for the expression level of the protein that covers the defined region on the gel. Spot volumes are still calculated by the 2D-gel image analysis software by summing up the per pixel color intensity measured on the gel.

5. Results and Discussion

5.1. Evaluation of the similarity measure

To test our similarity algorithm we constructed a set of spectra of which the underlying protein was well known. The set contained 558 spectra. We tested the performance of the algorithm by pair–wise comparisons of spectra in the test set. The resulting similarity values were binned either to the true correlation bin when the compared spectra had identical identification or to the false correlation bin, when the two spectra had a different identification.

We already stated that the two correlations should be equally weighted because of their orthogonality to each other. This is proven by the weighting scheme described in section 4.1. We have evaluated the weighting scheme by setting various values for the weighting factor *k*. A small value for *k* weights the mass correlation more than the rank correlation and on the other hand, a larger value of *k* only considers the rank correlation. The outcome of setting *k* equals to zero is shown on Figure 8.

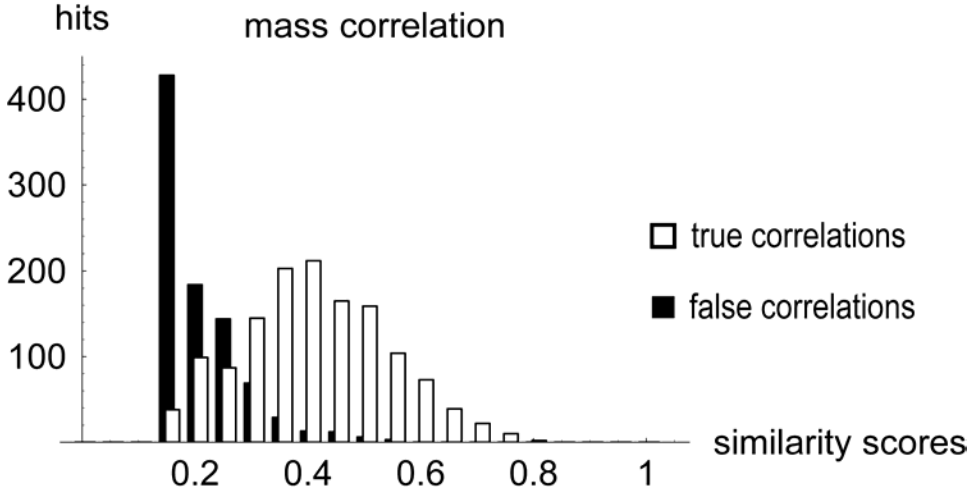


Figure 8: Pair-wise comparisons were carried out using the mass correlation alone as described in Chapter 4.1 without including the amplitudes of peaks. A significant overlap between the two bins is observed and false correlations are scored with similarity values up to 0.6.

When assigning *k* large values (*k >> 1*) only the rank correlation is taken into account. The evaluation of this procedure is shown on Figure 9. For better performance, if two peaks have intensities that are within 10% of each other, the same rank (tie) is assigned in our procedure.

As it is seen on Figure 10, the optimal weighting is achieved when setting *k* to 1. A *k* value equal to 1 means that both correlations are equally weighted. The weighting formula evaluates to form the geometric mean of the two correlations as it is described in equation (2) of section 4.1.2.

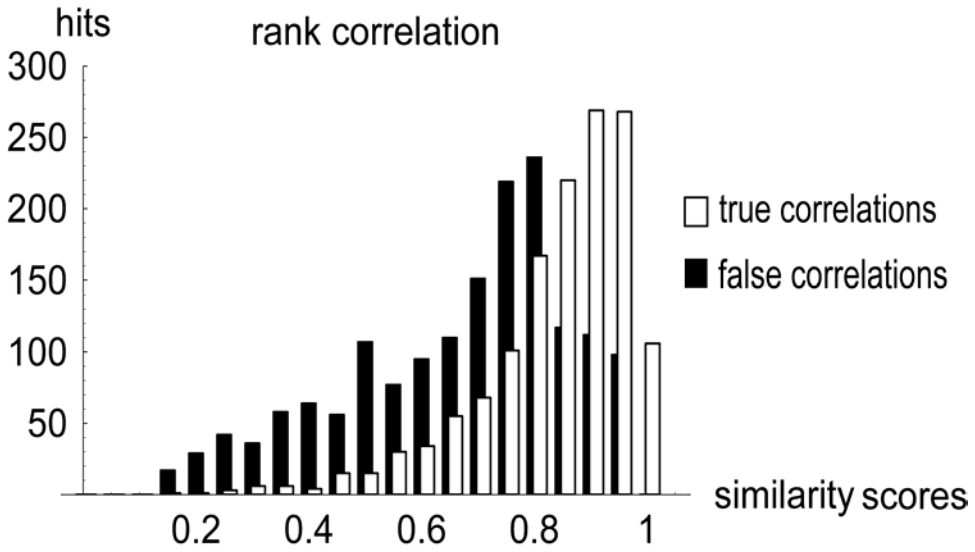


Figure 9: Pair-wise spectra comparisons of the spectra set were scored using Spearman rank correlation coefficient alone. The weighting factor *k* is set to a very large value *k >> 1*. The separating capability of this method is even lower than the method using mass correlation (see Figure 8).

As shown on Figure 10, using the combination of the two measures as similarity criterion achieves almost complete separation. Mass correlation alone, as seen on Figure 8, shows significant overlaps between false and true correlations. The same is true for pair wise spectra comparisons carried out using the rank order correlation alone as it is shown on Figure 9.

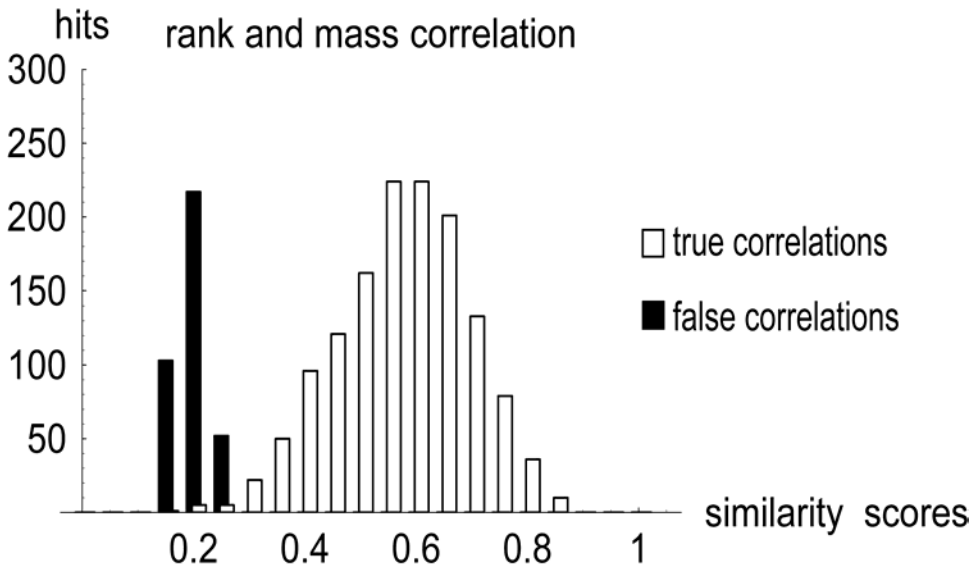


Figure 10: Combining mass and rank correlation of MALDI mass spectra allows almost complete discrimination between truly related spectra and unrelated ones in this sample of 558 spectra. The difference between the median false correlation bin and its true correlation counterpart is around 0.4

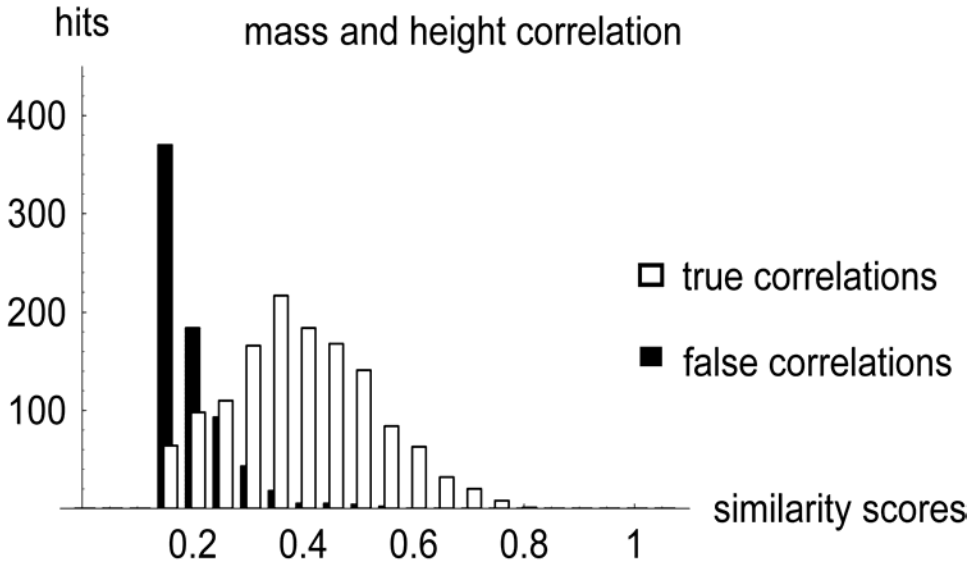


Figure 11: Pair wise spectra comparisons using mass spectra with intensities normalized to unit vector [36,48] show a significant overlap of scores for true and false correlations for the same spectra sample as in Figure 10. The difference between the median false correlation bin and its true correlation counterpart is only 0.2. While calculating the rotational angle between vectors in mass space is the usual approach to construct mass spectra libraries, it is clearly not a satisfactory method for separating MALDI PMF spectra.

We have compared our algorithm to the known measure of normalized dot product [36,48,49]. In this formulation only the *k* highest peaks overlapping between two spectra are taken into account and normalized to the total intensity of overlapping peaks. As seen

on Figure 11, the normalized dot product solution does not work with the same accuracy as observed when using the mass correlation in combination with the rank order correlation coefficient. Many true correlations are scored with low correlation values, meaning that complete separation of the two bins is impossible to achieve.
MALDI spectra are accurately compared using the algorithm that includes a robust measure of the intensities. The combination of mass correlation and rank correlation does not show false correlations above a similarity score of 0.3, whereas mass correlation alone or correlations with normalized ion current yield false correlations with similarity values up to 0.6. On the other hand, our method is the only method tested that did not miss true correlations. While it is obvious that adding intensity information to a spectra comparison is complementing it with orthogonal information content, it is crucial to understand that the absolute height of a MALDI peak does not denote a robust measure of its quantity. MALDI data is quantitative only in comparison to the other components contained in the same spectrum. Rank order reflects just this fact, and it is robust in presence of noise or separate components.

5.2. Clustering evaluation

5.2.1. Performance of the tree-based clustering

The UPGMA-tree building routine was used to build up a tree from a list of pair–wise comparisons of MALDI spectra. The list of comparisons and its scores describes a trigonal matrix. The similarity measurements were performed with the new spectra similarity algorithm described in the previous chapter. Complete sets of rooted trees have been built as shown on Figure 12. Clusters are formed by cutting edges of the tree to generate separate sub–trees that in turn should represent similar spectra accounting for the same protein identification. The separation of clusters is automatically done using a threshold parameter that is chosen upon cluster consistency. The threshold parameter is set in such a manner that a reasonable number of clusters is obtained and a large quantity of the previously compared spectra is assigned to a cluster.

We employed both static thresholds and dynamic thresholds (using significance of differences from average distance in the data). Static thresholds employed on the structure of trees often separate clusters that belong together. On the other hand by setting the threshold too low, many spectra are assigned to the wrong clusters.

Non-static thresholds are more suited to overcome the problem of cluster separation. However, there the problem arises by the fact that assumptions on the tree structure have to be made that cannot be generalized. In general, it can be stated that a tree structure is not an adequate description of the relationship between the spectra. It is very hard to disjoin the tree into the correct clusters using a static threshold parameter and it is similar difficult to break the tree into clusters by employing a dynamic threshold parameter.

The problem described above arises due to the non-ultrametric and non-additive behavior of the distances between two clusters. The ultrametric condition implies when two clusters, C_i and C_j are amalgamated that the distances between any leaf in cluster C_i and any leaf in C_j are the same. This is not the case when clustering MALDI spectra using the algorithm described in chapter 4.1, as well as when using any other measure of correlation. Similarly, most other tree constructing algorithms would fail as most of them rely on additivity or ultrametry of the distance measure. Therefore, we have to devise a different method for the analysis of sets of similar spectra.

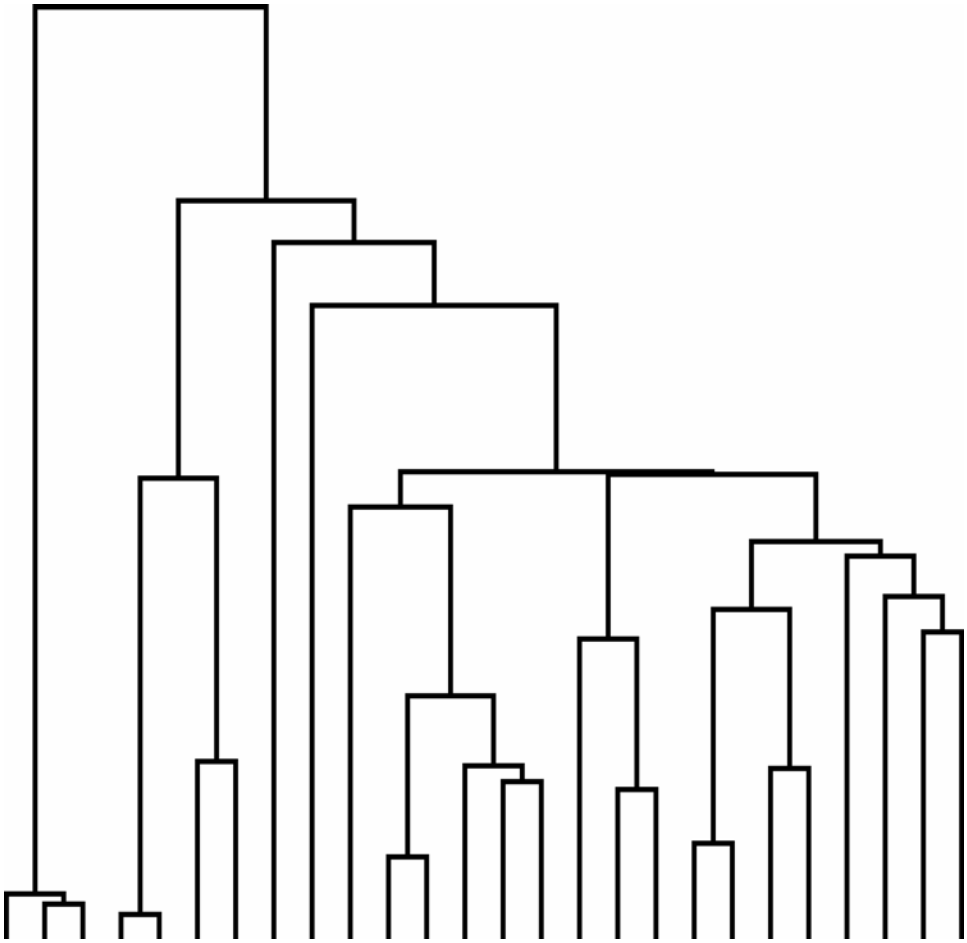


Figure 12: This figure shows a tree generated from pair wise comparisons of 26 spectra. In order to separate the tree in many sub-trees, a threshold parameter has to be defined. It can be demonstrated that many spectra fall out of clusters due to a threshold that is too harshly operating.

5.2.2. Performance of the graph based clustering

Again, correlations for each pair of spectra are calculated and a trigonal matrix is formed. The trigonal matrix serves as template to generate the list of highest similar pairs of spectra upon which the graphs have been constructed. Similarity graphs were generated by connecting spectra nodes to the nearest node whose edge has not been visited before.

As this method of constructing a graph relies only on the order property of correlations (stating that higher correlation values mean more similarity), it is robust against any of the factors that disturb the performance of tree building algorithms (e.g. noise, weakly linked spectra etc.). Moreover, the resulting graph is in general disconnected,

containing a forest of trees, each of them representing a cluster of similar MALDI spectra. An optimal iterative cluster consistency checking process can even improve performance. Cluster consistency is checked by comparing the number of total spectra in the data set to the number of spectra that have been assigned to clusters and by comparing the total number of different identifications from the data set to the number of clusters. A simple check of the selectivity and robustness of our method has been carried out by clustering two datasets from two different organisms (data not shown). Not even a single spectrum from one dataset appears in a cluster of spectra from the other dataset.

In order to increase performance and accuracy we checked that it is safe to ignore spectra comparisons whose number of overlapping peaks is lower than 4. This reduces the trigonal matrix to a sparse trigonal matrix. This reduction has positive implications on the scalability and therefore performance of the clustering algorithm, since only spectra are compared that have the potential to yield a high similarity score.

An example of such a sub graph is shown on Figure 13. This Figure shows an energy-minimized drawing of a graph of 38 *alpha-enolase* spectra. For purposes of drawing, an energy minimization is calculated using a 2D spring embedding algorithm. The idea of spring embedders is to simulate the graph as a system of mass particles. The nodes are the mass particles and the edges are springs between the particles. The algorithm tries to minimize the energy of this physical system[50]. As pointed out above, distances obtained by correlation methods do not follow easy metrics, so one should regard the drawing just as an illustration of the clusters.

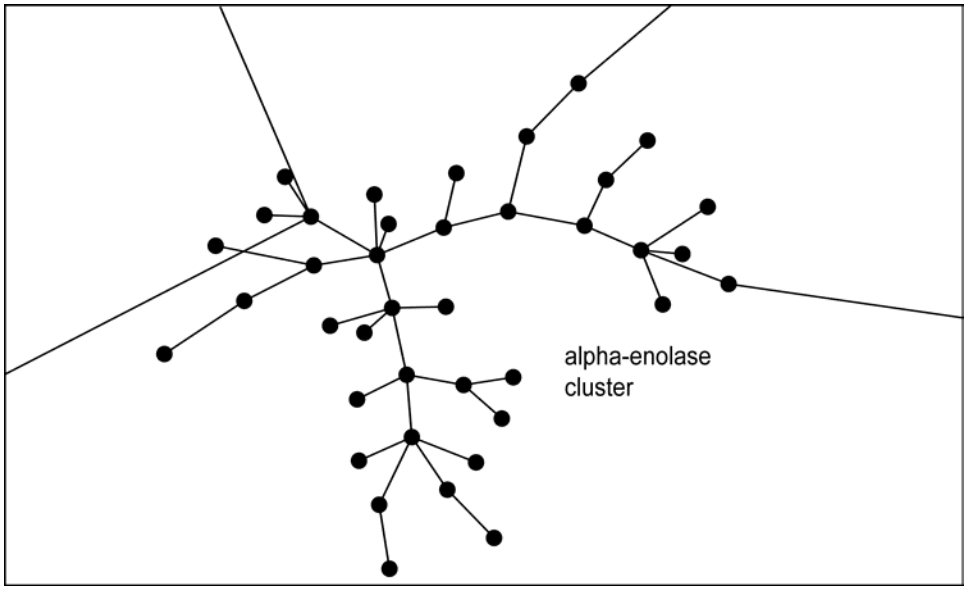


Figure 13: Graph of a single cluster of 38 spectra isolated from a set of ca. 1500 MALDI PMFs. The core region consists of spectra that have been identified as *alpha enolase*. Smaller distances correspond to stronger similarities. Note that average distances within the cluster are similar and very distinct from outbound connection distances. An iterative procedure has been implemented that employs this feature to separate self-consisting clusters.

As seen on Figure 13, the graph is not totally disconnected. There are a few outgoing connections to spectra at much larger distance than within the cluster. We have implemented a simple filter that removes inconsistent distances from clusters in order to obtain maximum cluster consistency.

The graph-clustering algorithm is much better performing than the tree-based clustering algorithm. As the threshold parameter in the case of graph clustering only polishes already separated clusters, there are no false assignments due to inadequately chosen thresholds.

Due to the better performance of the graph-clustering algorithm, we used this algorithm as the method of choice for all analysis of MALDI spectra samples.

5.3. Consensus spectra including rank ordered peaks

For every resulting cluster, a consensus spectrum has been generated. The consensus spectrum contains not more than 50 of the most abundant peaks from the clustered spectra. All consensus spectra were built following the procedure described in section

4.3. In general, a good consensus spectrum can be derived for a cluster that consists of more than 10 spectra. In clusters where the number of assembled spectra is greater than 10, a significant improvement of the peak parameters has been observed. These improvements consist of more accurate monoisotopic masses and of a more robust rank order of the peaks.

We conducted database searches for every consensus spectrum extracted from clustering in order to identify the protein. Identified proteins were searched again against the protein database searching tryptic peptides that were not included into Peptide Mass Fingerprint search. In general, these peptides are peptides with more than one trypsin cleavage site within the sequence, so called missed cleavages. As already stated in section 3, the PMF search procedure we employed does not use missed cleavages per default. The peak peptide matching results in an average number of nine peptide hits per consensus spectrum. As our means of generating consensus spectra tries to include just the most abundant peaks, the number of peaks does not reflect the number of theoretical peptides. We have already mentioned that about 80% of the spectra from a sample set can be assigned to clusters. 60% of these clusters and their consensus spectra are identified as containing peptides from at least one protein. Assemblies of protein mixture spectra result in consensus spectra that account for different kinds of peptides meaning that two or more identifications result from the peptide mass fingerprinting and therefore result in an increased number of peptide matches. However, in general, it can be stated that consensus spectra lead to safer and more robust protein identifications.

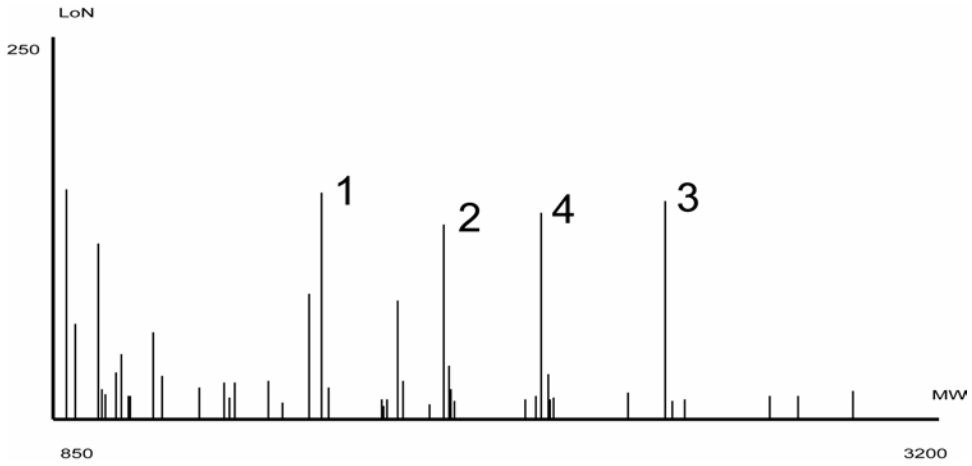


Figure 14: Shows a graphical representation of a consensus spectrum with its rank ordered peaks. As it is illustrated here, peaks with the highest averaged level over noise are not necessarily ranked better than peaks with a smaller level over noise.

A typical consensus spectrum derived from a cluster containing 14 spectra of the same kind is shown graphically on Figure 14. A table of the peptide identifications of the peaks assembled in such a consensus spectrum is shown on Table 2. As it is seen on this table, peaks with the highest level over noise are not necessarily ranked first because we assign ranks according to the prevalence of the peptides but not their ion intensity. It is also observed that peak peptide matches are prevalent in the first third of the consensus spectrum. Apparently, noise components or minor peaks are much less reproducible than well ionizing peptides from a protein digest.

In order to extract useful information in terms of peptide sequences and peak orders, we stored all data derived from consensus spectra in a database system.

Mass	LoN	dMass	Rank	pOccurrence	Sequence
2134.11	49.8793	1.79E-02	1	1	EAQTSFLHLGYLPNQLFR
1685.74	52.6356	2.18E-02	2	0.92857099	GSAFAIGSDGLCCQSR
2372.18	43.9949	2.14E-02	3	0.85714298	FLVLDEADGLLSQGYSDFINR
1111.68	36.4212	2.71E-02	4	0.92857099	ELLIIGGVAAR
2008.88	38.3587	0.025856	5	0.78571397	GKHYYEVSCHDQGLCR
2876.5	42.7084	0.022455	6	0.78571397	GIDIHGVPYVINVTLPDEKQNYVHR
1572.69	41.6644	2.47E-02	7	0.78571397	GHQFSCVCLHGDR
1153.56	25.2914	1.96E-02	8	0.78571397	MDQAIIFCR
1823.77	23.9341	0.039297	9	0.78571397	HYYEVSCHDQGLCR
2079.1	20.3296	3.37E-02	10	0.92857099	GIDIHGVPYVINVTLPDEK
1968.03	18.2592	0.022528	11	1	DQLSVLENGVDIVVGTPGR
1260.67	19.7872	0.035759	12	0.78571397	NQALFPACVLK
2138.17	18.2207	2.15E-02	13	0.92857099	ALIVEPSRELAEQTLNNIK
2572.32	28.8193	0.011019	14	0.35714301	
2595.37	18.8334	1.96E-02	15	0.64285702	
1200.57	20.8842	3.96E-02	16	0.71428603	EVKEWHGCR
1570.86	24.9929	0.034981	17	0.71428603	
2707.4	18.6234	0.031707	18	0.57142901	
1169.54	17.608	0.028671	19	0.78571397	
2669.49	17.2591	0.029805	20	0.78571397	
1246.67	17.7148	4.16E-02	21	0.35714301	
881.286	43.5903	5.79E-02	22	0.142857	
975.485	18.8786	3.54E-02	23	0.85714298	TDRLWER
1278.66	31.0869	8.25E-03	24	0.142857	
1626.79	18.1647	2.69E-02	25	0.71428603	

Table 2: This table shows the top 25 peaks of a consensus spectrum. The cluster is assembled from 14 spectra. Only two peaks, ranked 1 and 11, occur in every spectrum. A few peaks have a relatively high level over noise but are only occurring in two of the spectra. The rank ordering procedure takes care that these peaks get low ranks, despite their high level over noise.

5.4. Reference library of consensus spectra

We clustered spectra from the data sets described in section 4.4. Around 80% of the spectra could be assembled in clusters by our clustering method. Clustering resulted in a set of 5167 clusters from which a consensus spectrum could be extracted. 1715 of these consensus spectra in turn led to 1827 protein identifications, of which 1248 were unique. 933 identifications occurred more than once. As our means of generating consensus spectra tries to include the most abundant peaks, the number of peaks in the consensus spectrum does not reflect the number of theoretical peptides in the digest. This is shown on Figure 15.

This fact is explained by the overall observed peptide coverage that is in the range of 20% in peptide mass fingerprints measured by MALDI mass spectrometers.

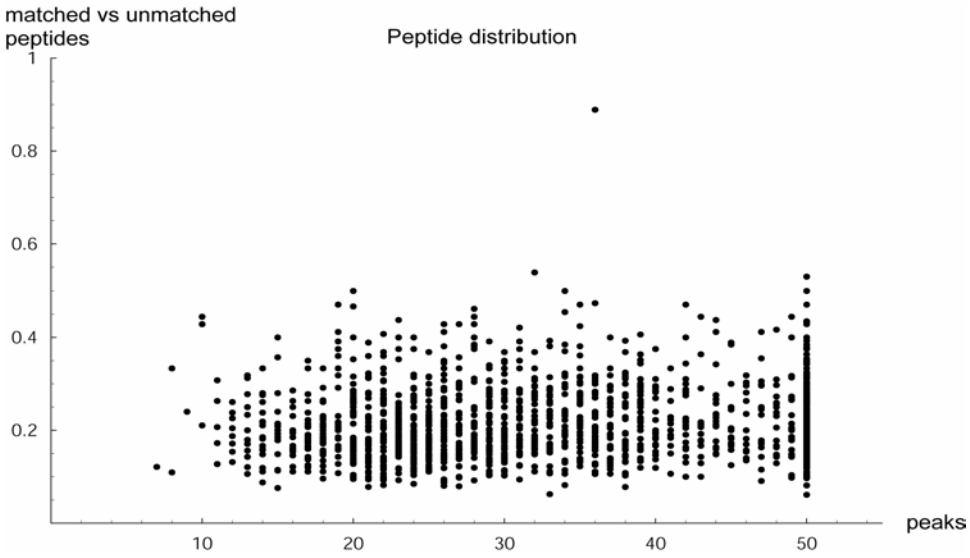


Figure 15: This figure shows how the peptides are distributed in consensus spectra and shows that the number of naturally occurring peptides does not correlate with the number of peptides in a consensus spectrum

Considering the completeness of sequence databases employed for searching, we would estimate that ~60% of the consensus spectra should yield an identification. Our identification rate is somewhat lower, a fact that is explained by the consideration, that only for large clusters (where the number of assembled spectra is greater than ~10) a significant improvement of the peak parameters can be expected. The peak–peptide mappings resulted in a total of 11235 unique peptide sequences, with an average of 9 peptides per consensus spectrum.

The reference library of consensus spectra serves for two purposes:

1) Instead of searching the peptide-database by peptide mass fingerprinting, we can compare the spectrum to be identified with the spectra in the consensus spectra database.

2) Consensus spectra can be used to elucidate peptide properties for peaks showing consistently high intensities.

5.5. Performance evaluation of the reference library

Identification of proteins using a library of consensus spectra has the big advantage of automatically considering only a realistic set of observable peptides. Considering that

the sequence coverage in MALDI PMF is typically in the range of 20%, this means that about 80%-90% of the usually considered peptide masses can be ignored without loss of search accuracy. As easily seen from the probabilistic mismatch estimate this will lead to several orders of magnitude differences in mismatch probabilities even for few matches and thus increase search sensitivity dramatically. The peptide mass fingerprinting algorithm has an observed identification rate that is between 40%-60% depending on the size of the sequence database.

Since we are able to cluster all the good spectra from a sample correctly (corresponding to 80% of the spectra), it is theoretically possible to identify 100% of the clustered spectra. This is because we employ the same strategy for clustering as for searching the reference library of consensus spectra. This would yield an increase in sensitivity of <60% to reach an identification rate close to 100%.

We carried out a simple test procedure to check the performance of such a reference library of consensus spectra on the one hand and to check the usefulness of our similarity algorithm for search the library on the other hand. The same test set was used as described in section 5.1 to test the accuracy of the algorithm. All the spectra of this data set of which the underlying protein is known were sent to the database of consensus spectra and compared to the spectra from the library. We only considered the most similar spectrum in order to assign identification to the spectrum. All the spectra that have a corresponding consensus spectrum in the library were correctly identified. The assignments failed only in cases were no consensus spectrum was available. In these cases no identification or a wrong identification was achieved. Additionally, we carried out the same test, using the mass correlation alone as the measure of similarity. By using the mass correlation alone, the rate of false identifications dramatically increased, not only in the case of a missing consensus spectrum (data not shown).

As mentioned before, a drawback of identifying a spectrum by searching the database of consensus spectra is that a consensus spectrum representing the searched protein has to be experimentally observed in previous measurements. Current progress in mass spectrometry instrumentation makes it likely that the task of establishing complete spectra libraries for commonly used laboratory organisms could be feasible in the near future.

5.6. Results of the peptide analysis

5.6.1. Missed cleavage patterns

We used previously established databases of consensus spectra from *Bacillus subtilis* and *Human* samples. With the obtained peak lists of consensus spectra we carried out database searches in order to identify underlying proteins without considering missed cleavages or posttranslational modifications. If a spectrum was identified, it was compared to the theoretical digest of the identified protein again, allowing this time two missed cleavages within the sequences. The latter procedure increases the amount of peak peptide matches.

In total, we have analyzed 3251 theoretical peptide sequences from *Bacillus subtilis* and 6859 theoretical sequences from *Human* proteins. These 10110 sequences contained 5760 missed cleavage sites, which were subjected to an analysis of the two cleavage site flanking amino acids. The results of this analysis are shown in Table 3.

Missed cleavage pattern	
Pattern 1	[WYF][RK][^RK] or [^RK][RK][WYF]
Pattern 2	[DE][RK][^RK] or [^RK] [RK][DE]
Pattern 3	[^RK][RK][RKH]
Pattern 4	[RK][P]

Table 3: This table shows the observed missed cleavage pattern. Pattern 1 is explained as follows: A missed cleavage pattern is detected when either W or Y or F is followed by R or K and the latter are not followed by neither R nor K. These patterns account for more than 90% of the missed cleavages.

As we pointed out above, consensus spectra can be used to elucidate constraints on the physico-chemical properties of observed peptides. This serves the purpose of directly reducing the set of peptides taken into consideration for identification purposes.

A dramatic reduction of the size of the peptide database can be achieved by obtaining information on missed cleavages and information on the distribution of K-/R-ending peptides. The four missed cleavage patterns listed in Table cover 91% of the most often occurring detected missed cleavages. Pattern four is well known because trypsin is unable to cleave when arginine or lysine is followed by a proline. Pattern two and part

of pattern three have been described in literature before[51]. Pattern one has not been described in literature before due to the fact that the three amino acids tryptophan, phenylalanine and tyrosine are low abundant amino acids. Part of pattern 3, the histidine occurring at the right side of a potential missed cleavage site is also new, likely because histidine is a low abundant amino acid.

5.6.2. Distribution of R- and K-ending peptides within the first ranks

The database of consensus spectra includes the rank order of each peak. We queried the database for the occurrence of lysine– or arginine– ending peptides within the first few ranks of the consensus spectra. The result of that investigation is shown on Figure 16.

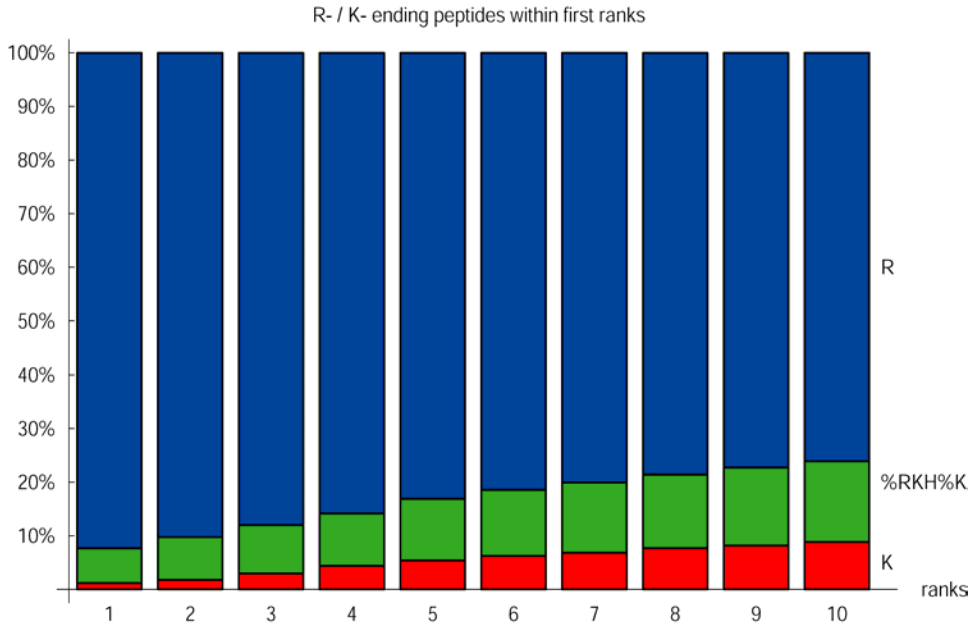


Figure 16: As this figure impressively demonstrates, there is practically no lysine ending peptides observed within the first then ranks. Some lysine ending peptides are observed but they are missed cleavages, carrying R, K or H within the sequence. The mayor part of peptides within the first ten ranks are R ending peptides.

As it is obvious from the data, arginine–ending peptides are much more frequent among the top-ranking peptides then their statistical expectancy, with the frequency of observing a K–ending peptide without another basic residue among the first 10 ranks being less than 8%. This holds even for multiple lysine residues in the sequence. As shown on Figure 17, the occurrence of multiple lysine residues in matched peptides

without arginines or histidines is distributed approximately like a binomial[52] with a frequency of 6.5% (instead of 27.8%, as expected by amino acid frequencies). Interestingly, the performance of R-residues is a property of MALDI. The most likely explanation for this is the higher basicity of the guandinium group in gas-phase[53], which leads to a non equal distribution of the protons after a short time in the relatively hot plasma of the extract.

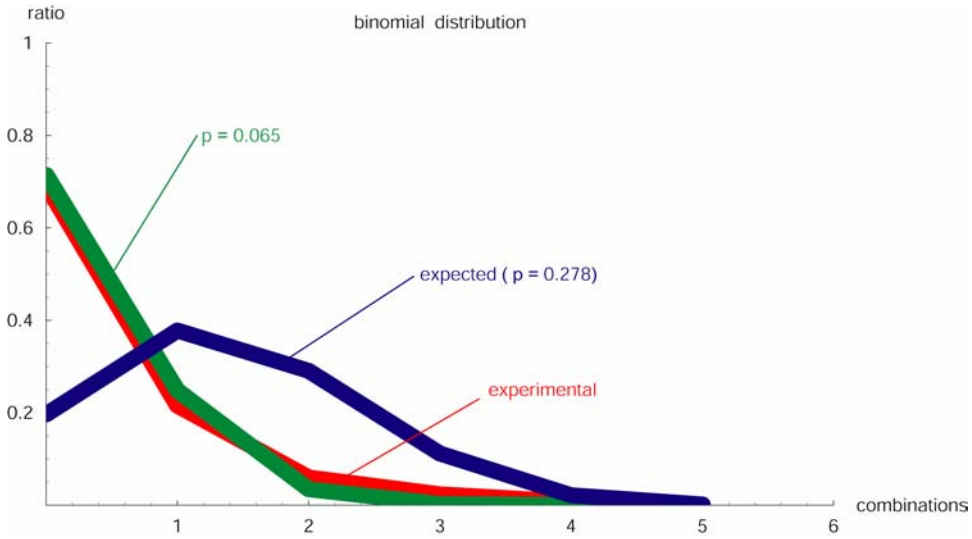


Figure 17: Binomial distribution of K-ending peptides not containing R or H. The experimentally determined values fit to the binomial distribution of values with a probability p of 0.065. Theoretical sequence database analysis predicts a 4.3 fold higher abundance of lysine–ending peptides.

5.6.3. Kyte and Doolittle hydrophobicity plot

Several papers presented hypotheses about an increased hydrophobic character of peptides that were observed in MALDI mass spectrometric measurements [54-57]. We have evaluated the hydrophobicity and the gradient of hydrophobicity for each peptide in our database. This was done using the per amino acid hydrophobicity score as defined by Kyte and Doolittle[42]. The outcome of this analysis is shown on Figure 18. No difference between matched and unmatched peptides could have been observed. Additionally, we calculated the hydrophobicity and the hydrophobicity gradient in dependency on peptide rank for the ten top-ranked peptides. We could not observe positional dependencies (amphiphilic character) for the hydrophobicity of amino acid side chains.

The observations reported in the papers were drawn on a small sample of peptides. It appears that the prevalence of small amino acids in the vicinity of the C-terminal residue (see below) was misinterpreted in terms of hydrophobic/amphiphilic character of the peptides.

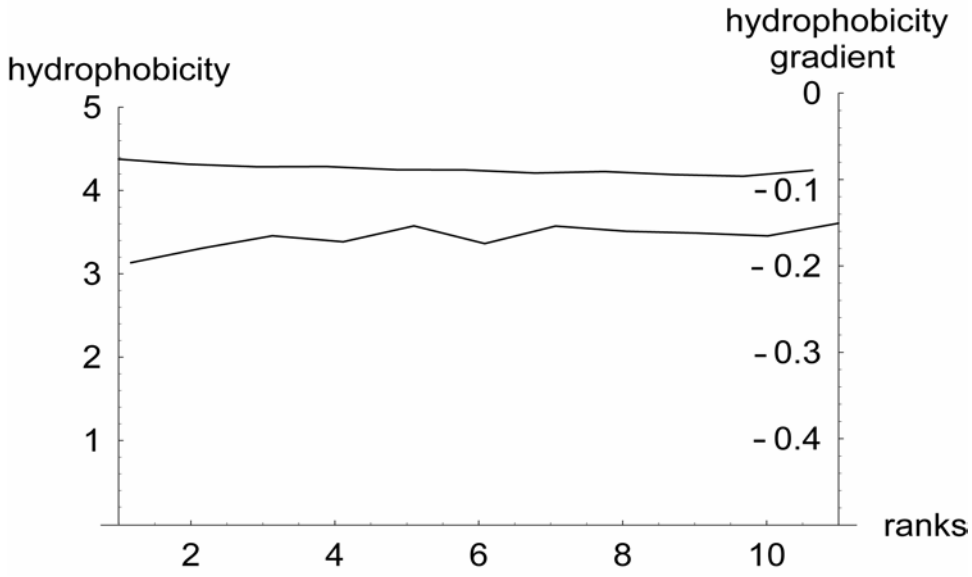


Figure 18: This figure shows two properties of matched peptides. It shows the hydrophobicity of peptides within the first ranked peptides and the gradient of hydrophobicity of each peptide ranked within the first ten peptides. The gradient is defined as the slope of hydrophobicity from the N- to the C- terminus of a peptide sequence.

5.6.4. Amino acid distributions as indicators of peptide ionization characteristics

For the calculation of the relative entropy per position and amino acid, we took into account 10 amino acids on the C-terminal site of the peptides. We assumed that the N-terminus features a similar behavior as Lysine in the ionization process, so that both N-terminal amino acids as well as Lysine terminated peptides can be ignored. We analyzed ten positions before the proton accepting Arginine, which is assumed to be sufficient to reveal possible differences in amino acid distributions by comparing matched to unmatched peptides. 8500 peptides containing at least 10 amino acid residues have been aligned starting at the C-terminus. Entropies were calculated by taking the frequencies of amino acids at every position in the alignment. The result of this analysis is shown on Figures 19 and 20. We observed significant differences in

residue distribution patterns between matched and unmatched peptides. Peptides detected in consensus MALDI MS spectra contain a marked excess of small amino acids (alanine, glycine, valine), while the occurrence of acidic amino acid residues in immediate neighborhood of the C–terminal basic amino acid is reduced.

On the other hand, analysis of the relative entropies of amino acid positions in peptides revealed clear differences between matched and unmatched peptides. Whereas the unmatched peptides mostly carry acidic amino acids before the ion acceptor, the matched peptides prefer small amino acids like gylcine, valine and alanine at this position. These three amino acids are more prevalent at all ten positions before the C-terminal arginine or lysine. Only at position four, five and eight the relative entropies of D and E are higher than the ones for A, G or V. Such a pattern is not observed in the unmatched peptides, where only D and E residues contain information relative to the background distribution, probably reflecting a general prevalence of ion bridges in protein secondary structures.

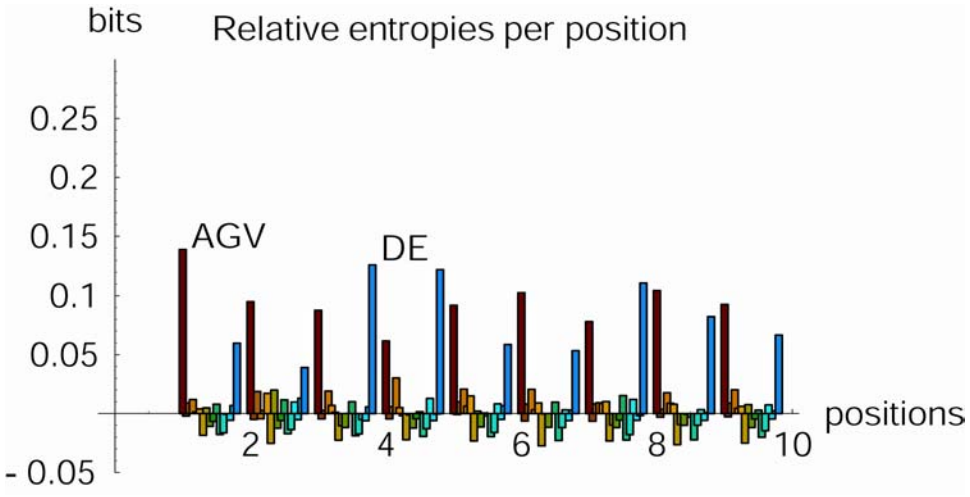


Figure 19: Information content of amino acid residues at the C-terminus of peptides detectable in MALDI PMFs. Entropies for arginine and lysine have been removed because their high information content is caused by the cleavage specificity of trypsin and suppresses the signal of other positions. Significant differences from the natural amino acid distribution could be detected for small amino acid residues (alanine, glycine, valine) and for acidic residues (glutamic acid, aspartic acid).

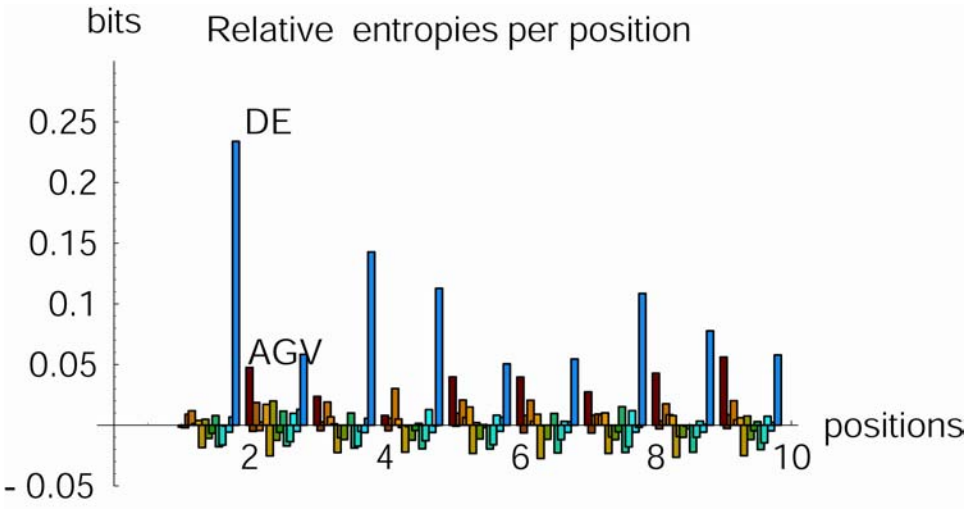


Figure 20: Information content of amino acid residues at the C-terminus of peptides not detected in MALDI PMFs. Significant differences from the natural amino acid distribution could be detected for acidic residues (glutamic acid, aspartic acid), but, in opposite to the data presented on Figure 19 not for small amino acid residues (alanine, glycine, valine). As discussed below, this could be interpreted in terms of different structural requirements for ionization in MALDI experiments.

Therefore, peptides with high ionization potential in MALDI have a marked prevalence of small amino acids over the background. While it is difficult to interpret this observation in terms of peptide structures or solution behavior, it could be speculated that the increased content of small amino acids provides a structural flexibility that aids the crystallization of the peptide with the matrix substance.

The knowledge about peptide properties gained from these studies has been directly implemented in the Peptide Mass Fingerprint algorithm described in chapter 3.3. Addition of this information improved the search procedure to achieve a false positive rate of 0.25% (down from ca. 5%) at the same level of sensitivity. This was tested using a database whose theoretical peptide masses had been shifted by 3 Da.

5.7. Noise free consensus spectra

By using the filtering techniques described in chapter 4.6, we were able to clean the datasets from noise. Known peaks as internal standards, enzyme derived autolysis peaks and unwanted contamination (e.g. kreatine peaks) are excluded from the spectra

after successful peak annotation. Polymer peaks also are eliminated before the dataset is further processed.

5.7.1. Elimination of polymer peaks

By using the autocorrelation function described in chapter 4.1 (Formula 3), we are able to search for polymers that have a monomer mass ranging from 1 to ~100. We set the endpoint of the range to 100, assuming that polymers with larger monomer masses would not yield enough of the polymer peak pattern to unambiguously identify the polymer. Peaks that are larger than 1/10 of the total ion flux at lag zero are potential polymer peaks. These peaks have to occur at several times within a monomer specific mass frequency.

An evaluation of the autocorrelation of a polymer-containing spectrum is shown on Figure 21. To clean a spectrum from such polymers, it has to be searched for peaks that have a difference in mass that corresponds exactly to the polymer unit mass. This is done using an adjacency list.

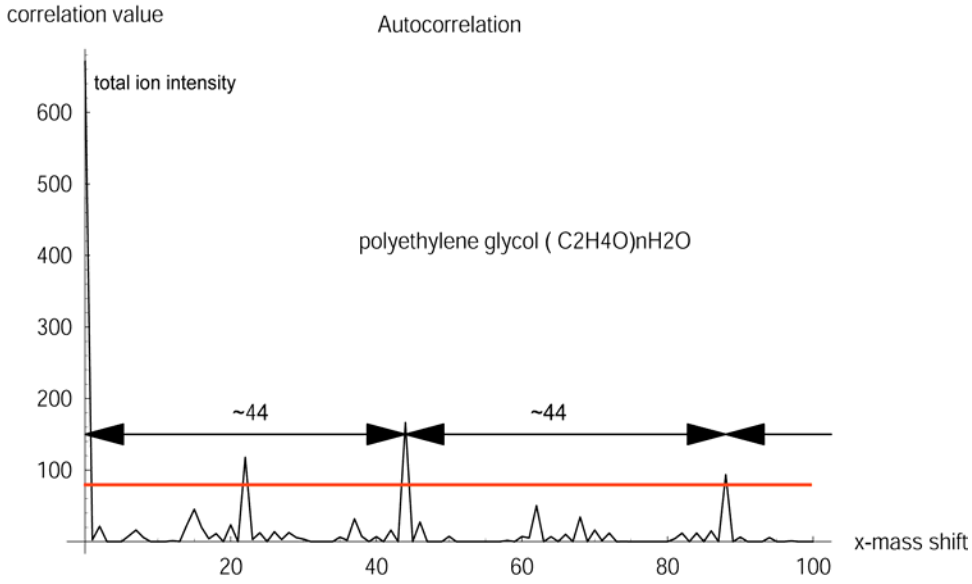


Figure 21: This figure shows the autocorrelation of a polymer contaminated spectrum at different mass-shifts x. The spectrum used here contains polyethylene glycol, a polymer whose monomers each have a molecular weight of ~44. The red line marks the cutoff value for peak detection.

5.7.2. Elimination of noise peaks using spot locations on 2D-gels

Using the spectra information gained from clustering, we can calculate the area that is occupied by the spots of which the spectra have been derived from. Every spectrum that has been measured was automatically excised from a 2D gel. The actual location of the spot is stored in a database along with the spectral information. This means for every spectrum we know its gel coordinates. By calculating the area that the envelope of the point set takes on the gel, we are able to decide whether a cluster is a noise cluster or not. Figure 24 shows the spot distribution of noise cluster and that of a protein cluster.

We use two criteria to decide whether a cluster is derived from spectra that contain noise peaks or not. As a first criterion, we state that spots located in an area that is larger than 15% of the whole gel must contain noise. This condition is sometimes not sufficient because the same protein can occur several times on a gel in modified form or as fragment. Consider the case of a protein that occurs in a full length and a proteolytically cleaved form. The majority of peptides and therefore peaks remain unchanged for both forms. The two types of spectra would fall into the same cluster, even though the spots that are the source of the spectra are located at completely different places on the gel. The spot envelope covers an area larger than 15%, even though the spectra are not noisy. Therefore, we introduced a second criterion that treats such a case differently. We calculated for every sub-species the area it covers. These areas should still not be larger than 15% of the gel. Secondly, we checked whether these areas contain at least 40% of the spectra. Both conditions have to be fulfilled to pass the noise test.

If a cluster is characterized as noise cluster (see Figure 22), a consensus spectrum can be generated. Such a noise consensus spectrum is shown on Figure 23. As it is seen on the spectrum image, there are many putative matrix peaks in the low molecular weight range. For each detected noise cluster and its consensus spectrum, its five most abundant peaks are declared as noise peaks and excluded from the whole dataset. As the dataset is freed from the most prominent noise peaks, another round of clustering leads to higher clustering accuracy. This can be done iteratively, collecting and

excluding the noise peaks from the dataset after each round of clustering until no more noise is detected.

As the clustering is improved by deleting the most prominent noise peaks at every clustering cycle, the clusters become increasingly self-consistent and accurate.

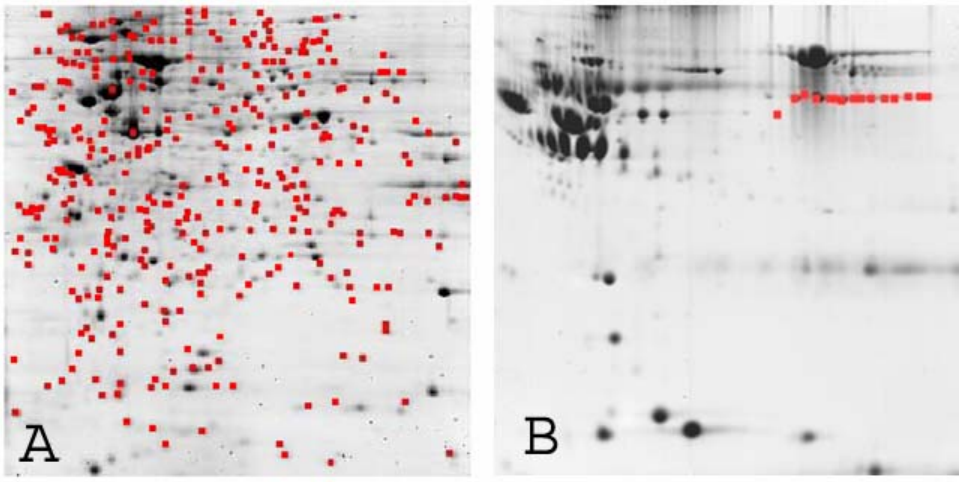


Figure 22: This figure illustrates a noise cluster (gel A) and how the spots are uniformly distributed all over the gel. The area covered by these spots easily exceeds 15% of the gel. Gel B however, shows the spot distribution of a cluster whose spectra are noise free.

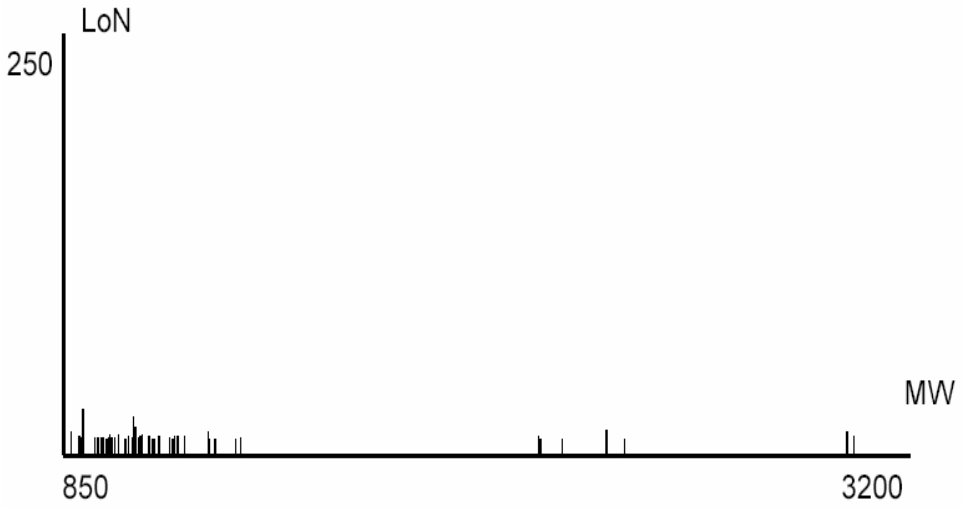


Figure 23: This figure shows a consensus spectrum generated from a cluster whose spectra are assembled together due to noise peaks. Matrix peaks and chemical noise peaks are typically observed in the range of 850-1050 MW.

5.8. Gel matching using spectral information

After the noise elimination resulting clusters are noise free and the cluster consistency is high. Thus, we can begin to map the cluster information to the original gels. This is done as described in section 4.7. We calculated a Voronoi diagram over the gel so that each spot is surrounded by a polygon that partitions the gel such that each spot is nearer to its central point of the polygon than to any other spot. A Voronoi diagram drawn on a gel is shown on Figure 24. Having declared an area per spot, for illustration purposes we colored all the polygons belonging to the same cluster with the same color. This leads to a raw picture of the gel and its clustered spots as shown on the right side of Figure 24. As it can be seen on the Figure, many spots could not be assigned to a cluster. This is because in routine operation, there are spectra where a proper measurement was not possible, or even where a spot could not have been picked.

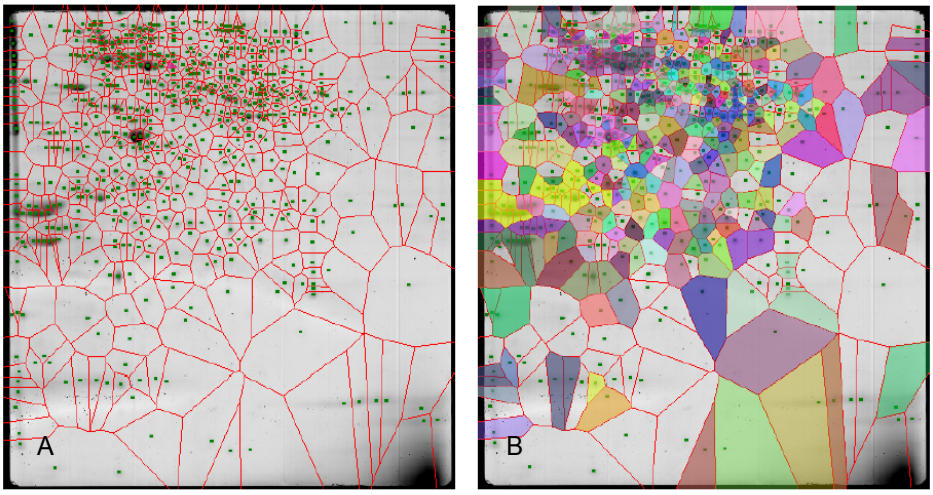


Figure 24: The left side of this figure shows a Voronoi diagram drawn on top of a gel image. As it can be seen, every single spot has its own polygon-defined region. These polygons are filled with colors, every color representing a separate cluster. This information is extracted from the spectra assemblies.

As it desirable not to neglect these spots for quantification purposes, we used the methods described in section 4.7 to add the information missing in the cluster in order to complete the picture.

The approach we employed uses the information from two or more gels from the same sample to add missing spot information. The result of this approach is shown on Figure 25.

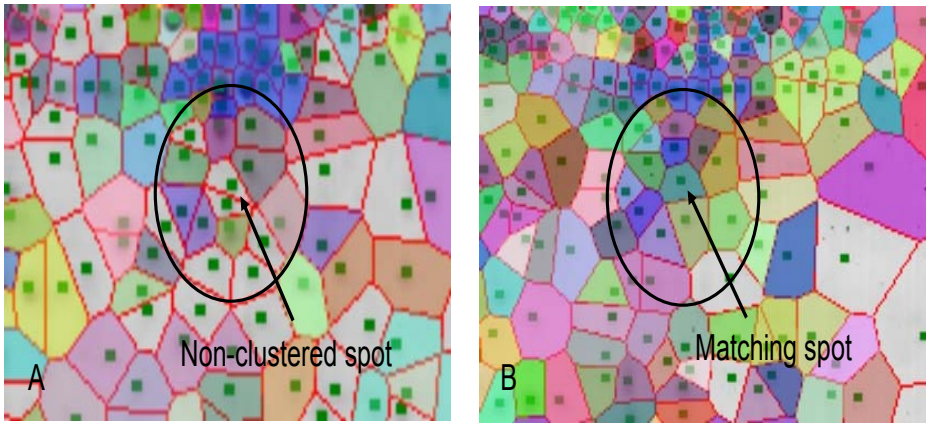


Figure 25: This figure illustrates the surround matching on other gels. The encircled regions denote the matching of surroundings on the left side to surroundings on the right side. The non-clustered spot in the center of the circle on the left side will be assigned to the same cluster as the green colored spot, seen on the center of the circle on the right side.

In that method, we calculate the 15 geometrically nearest neighbors of every spot on every gel in the sample set. These sample sets usually contain 2 or more gels of exactly the same sample (replicas). As this method is depending on replicas, the more gels there are the better the overall coverage of the protein species is in the end. Having the list of surrounding clusters for every spot, we can try to assign non-clustered spots to clusters. This is done by comparison of the surroundings of every spot to all surroundings of spots on other gels in order to match the surroundings of the non-clustered spots. We assign a non-clustered spot to a cluster if 35% of the surroundings on another gel overlap and the center of the surrounding is not further away than ±2% in x-y coordinates. The non-clustered spot is assigned to the cluster to which the center-spot of the matching surround spot list belongs on the other gel. This procedure is repeated for all the non-clustered spots.

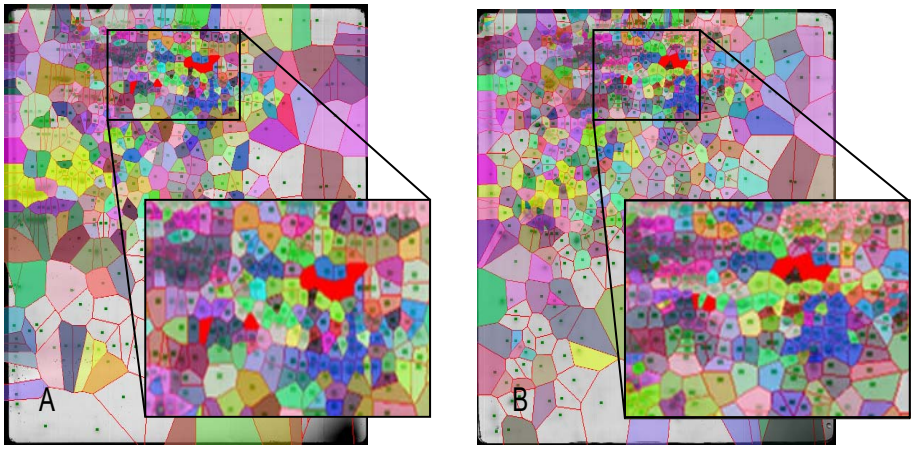


Figure 26: This figure shows the highlighting of two regions on different gels. These regions have been derived from clustering information and from the addition of missing spot information. The consensus spectrum is identified as *chaperonin containing tcp1*.

In general, the clustering procedure covers about 60 to 70% of the gels. The spot assignment method increases the coverage to reach 75% and finally another increase in 5 to 10%, depending on the overall quality and clusterability of the sample set, is achieved when adding spots due to neighborhood comparisons.

Using the spectra information from clustering, we were able to establish a method to compare gels allowing the comparisons of gels without comparing spots on the level of identification nor on the level of their location on the gels.

Using only counts of protein identifications tends to miss significant differences because there are many spots whose spectra are not identified. This situation does not occur when employing our method due to the fact, that a spectrum does not need to be identified to be clustered with similar spectra. Our approach is also independent from gel image comparisons in order to assign spots to specific clusters. The gain in reliability here is that we use regions on the gel that have the same cluster identity and not their spot image boundaries to assign spots to clusters.

As seen on Figure 27, we can match the same region on two very different gels from the same sample, which would have been difficult using the gel matching methods.

5.9. Accurate comparisons of gel sets

We established a method to accurately compare large sets of gels and mass spectrometric data measured from it on the level of mass spectrometric information. For the first time, our method allows a robust statement about protein expression differences when comparing two or more large different datasets.

In order to compare the gel sets, they have to be clustered separately as described in chapter 4.2. Clustering results in a set of clusters that contain around 80% of the spectra. This number corresponds to the estimate of acceptable spectra per sample set. The obtained clusters are then mapped to the 2D-gels. The resulting regions, each defining a separate entity, cover about 80 to 90% of the gels. Once the clusters and therefore the regions are assigned to their location by gel cluster mapping, the sum of the individual spot volumes can be calculated. After completion of these procedures, a comparison of all consensus spectra is carried out. The consensus spectra are compared pair–wise using the similarity measure described in section 4.1. As the consensus spectra are supposed to be noise free, the correlation value is thought to be significant even if it is below a similarity value of 0.3.

It is possible that a consensus spectrum from one gel set matches more than one consensus spectrum from the other gel set. That is the reason we decided to take only the most similar consensus spectrum for further processing. For every similar pair of consensus spectra, the change in overall spot intensity is calculated as follows: for a pair of spectra $i \rightarrow j$ the change in intensity is $\log 10\left(\frac{V_i}{V_j}\right)$. Thus, a change in intensity of 1 corresponds to a ten fold increased expression of protein *i* in comparison to protein *j*. V_i and V_j respectively denote the accumulated spot volumes derived from the spots contained in a region, which in turn is defined by consensus spectrum *i* and *j* respectively.

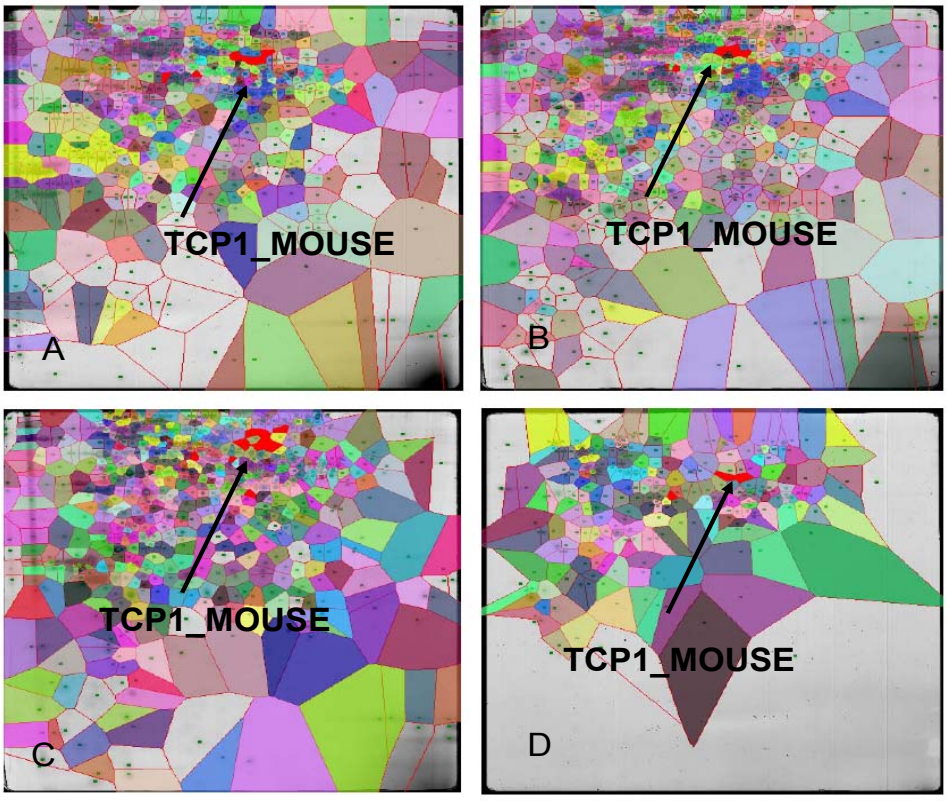


Figure 27: Highlights the region deterministic for *chaperonin containing tcp1*. The upper two gel images are from the same gel set whereas the lower two gels belong to another gel set. The similarity of these two regions was derived using the spectra comparison algorithm described in Chapter 4.1.

The comparison results in two categories of consensus spectra. The first category describes consensus spectra that have a consensus spectrum in the other gel set that is similar to it. For these pairs of spectra, we can calculate the change in intensity, which is an estimate of the difference in protein expression of that particular protein. The second category describes consensus spectra that have no similar counterpart in either of the gel sets. For these consensus spectra, a calculation of the change in intensity is not possible. However, if the presence/absence can be confirmed, these spots provide valuable information about protein expression differences larger than the dynamic range of the detection method.

The results of a gel set comparison can also be illustrated graphically as it is shown on Figure 27. The same region, accounting for the same protein, is highlighted on gels belonging to two different gel sets.

We are now able to match gels on the level of spectra and draw conclusions that are more robust than before. As we are comparing gel sets on the level of spectra content, we obtain expression level differences even if identification cannot be achieved (e.g. if no sequence database is available for an organism). Obtained consensus spectra can be selected and further analyzed using different identification methods.

5.9.1. Performance evaluation of the new method

A study entitled "Differential Expression Profiling of Human Pancreatic Adenocarcinoma and Healthy Pancreatic Tissue" from our lab has been published by *Lu et al*.[58].

This study has been carried out using the traditional method of analysis of the differentially expressed proteins. Differences in spot intensities were evaluated by 2D gel analysis software (Bio-Rad PDQUEST 6.2). Individual spot volumes have been estimated. The spots from two master gels, one per sample, have been analyzed and identified. These two gels served as templates for the remaining gels to assign protein identifications to spots that have been matched by the gel comparison software.

In this study, about 70 proteins were found in pancreatic carcinoma tissue that are 2 or more fold expressed in comparison to the control. It took an experienced operator several weeks to derive this information.

Since the samples have been analyzed in our proteomics laboratory, the data including 2D gel images and mass spectra were available to reproduce the results. We employed our newly established method to re-analyze protein expression differences.

We clustered spectra from each gel/spectra set separately to generate protein specific cluster assemblies as described in chapter 4.2 and extracted a consensus spectrum for every cluster as described in chapter 4.3. Regions which are defined by the spectra clusters were mapped onto corresponding gels in order to establish a clustered gel picture. Consensus spectra were then identified via traditional PMF or via the reference library of consensus spectra. In order to perform gel matching, all consensus spectra derived from clustering of the carcinoma spectra set were compared to consensus spectra from the control spectra set.

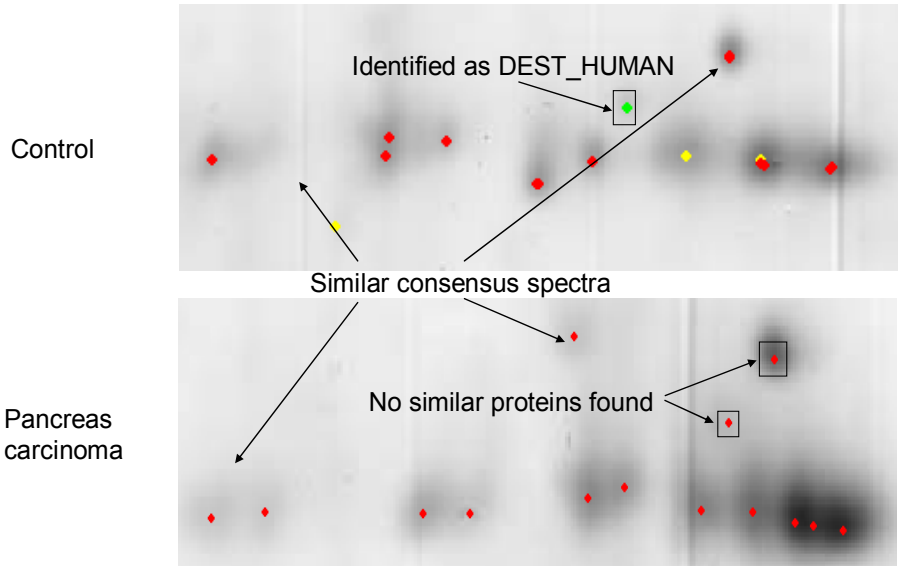


Figure 28: DEST_HUMAN (Destrin (actin-depolymerizing factor) (adf)) has been identified only in the control. It was suggested by *Lu et al.* that this protein is two fold overexpressed in the carcinoma sample. In the region of Destrin no similar spot or spectrum was detected that shows a two times higher expression level.

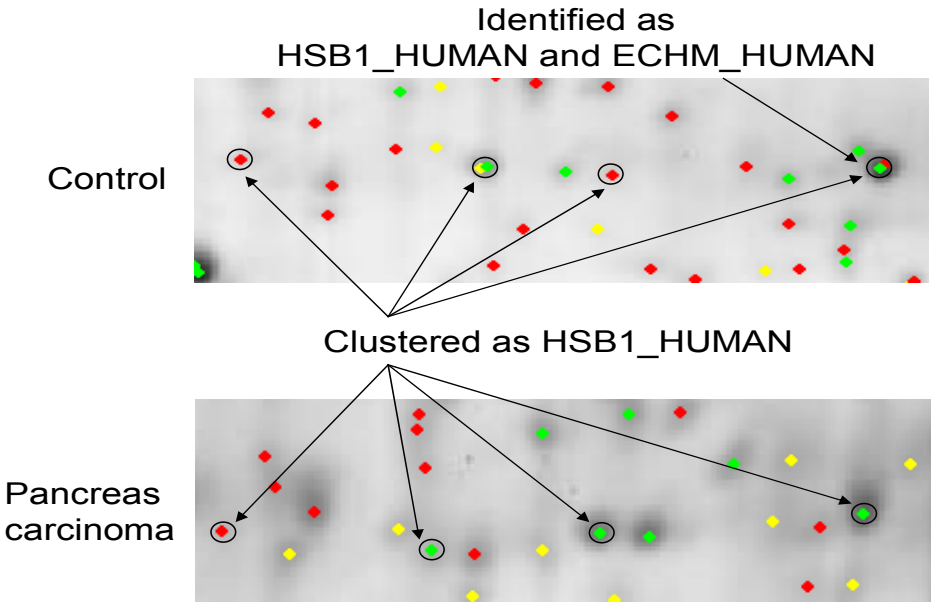


Figure 29: ECHM_HUMAN (enoyl-coa hydratase, mitochondrial precursor (ec 4.2.1.17)) was identified only in the control sample. Five spectra were measured from this spot, three of them identified as HSB1_HUMAN (Heat shock protein beta-1) and two of them as ECHM_HUMAN. However, as it is more likely that the major protein of this mixture spot is HSB1_HUMAN, ECHM_HUMAN might not be overexpressed in the carcinoma sample as described by Lu et al. All the spectra from the marked spots were clustered together and the resulting consensus spectra have been identified as HSB1_HUMAN.

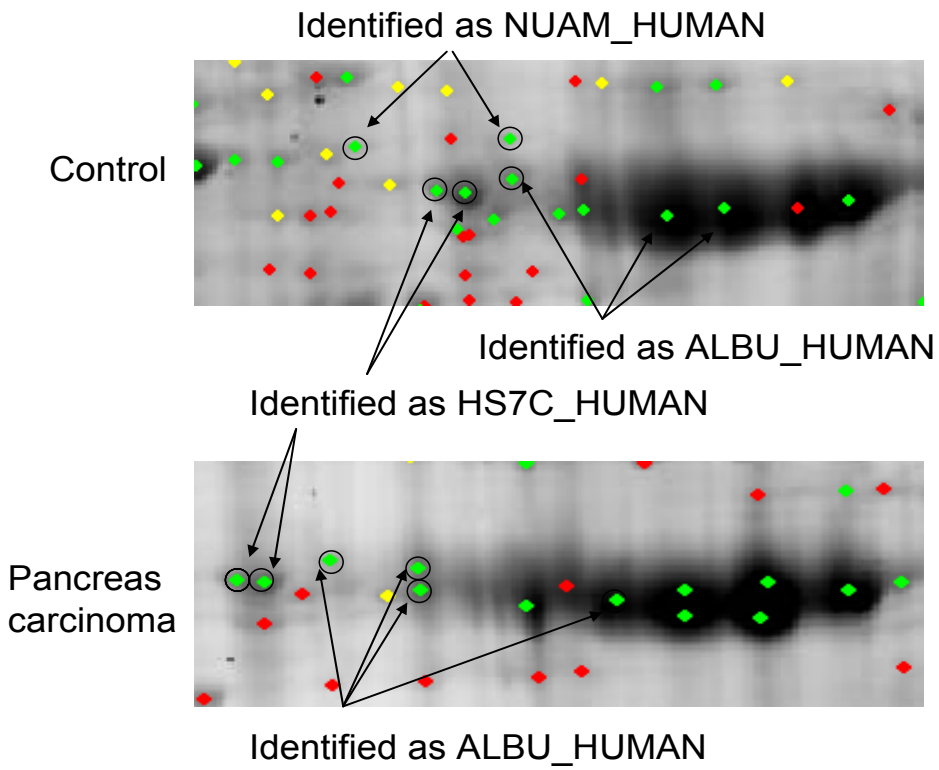


Figure 30: NUAM_HUMAN (Nadh-ubiquinone oxidoreductase 75 kda subunit), which is described as overexpressed in the pancreas carcinoma sample was identified only in the control. This figure shows similar regions on two gels, the first is a gel from the control and the second gel is from the pancreas carcinoma sample. Similar regions are highlighted on the two gels. There is no spot observed in the NUAM_HUMAN region that justified an assumption of NUAM_HUMAN being two fold overexpressed in the carcinoma sample.

We compared our expression analysis results to the list of overexpressed proteins described by *Lu et al*. 43 proteins of the 70 overexpressed proteins could be detected. However, for 27 proteins given in their list, we could not observe expression differences. A list of these proteins is given in Table 4. The table contains several proteins of which a different isoform has been detected. It also contains proteins that are clearly false positive identifications, which would be enriched in any gel-comparison analysis. Since we were able to reduce the false positive success rate of our peptide mass fingerprint algorithm to a minimum, we believe that the protein isoforms we detected using our consensus spectra are the correct ones.

Additionally, the list of differentially expressed proteins obtained from gel image comparison contains a few clear artifacts of the image comparison process. Possible explanations for these misassignments are given for three of these proteins (DEST_HUMAN, ECHM_HUMAN and NUAM_HUMAN) on Figures 28-30 in

detail. As our method requires substantial spectral information, it is not vulnerable to mismatches of image analysis that occur in the vicinity of large protein spots or in areas of insufficient gel quality.

Table 5 describes a list of proteins expressed only in the carcinoma tissue.

Protein	AccNo	Description	Status
humangp:CHR7-FSC2	O14926	sw:fsc2_human: fascin 2 (retinal fascin).	different isoform detected
Sw:CALD_HUMAN	Q05682	caldesmon (cdm).	not found in datasets
Sw:COF1_HUMAN	P23528	cofilin, non-muscle isoform (p18).	not found in datasets
Sw:DEST_HUMAN	P18282	destrin (actin-depolymerizing factor) (adf).	not found in carcinoma sample
Sw:AMPL_HUMAN	P28838	cytosol aminopeptidase (ec 3.4.11.1)	not found in carcinoma sample
Sw:EL3A_HUMAN	P09093	elastase iiia precursor (ec 3.4.21.70)	not found in datasets
Sw:APE_HUMAN	P02649	apolipoprotein e precursor (apo-e).	not found in datasets
Sw:CALX_HUMAN	P27824	calnexin precursor(p90)	not found in datasets
Sw:CYPH_HUMAN	P05092	peptidyl-prolyl cis-trans isomerase a (ec 5.2.1.8)	not found in datasets
Sw:TCPG_HUMAN	P49368	t-complex protein 1, gamma subunit	not found in carcinoma sample
Sw:DLDH_HUMAN	P09622	dihydrolipoamide dehydrogenase (ec 1.8.1.4).	not found in carcinoma sample
Sw:NUAM_HUMAN	P28331	nadh-ubiquinone oxidoreductase 75 kda subunit	not found in carcinoma sample
Sw:ECHM_HUMAN	P30084	enoyl-coa hydratase, mitochondrial precursor (ec 4.2.1.17)	not found in carcinoma sample
Sw:SYW_HUMAN	P23381	tryptophanyl-trna synthetase (ec 6.1.1.2)	not found in datasets
Sw:HSBX_HUMAN	O14558	heat-shock 20 kda like-protein p20.	different isoform detected
Sw:IQG1_HUMAN	P46940	ras gtpase-activating-like protein iqgap1	only 1 good spectrum
Sw:PBEF_HUMAN	P43490	pre-b cell enhancing factor precursor.	not found in datasets
Sw:RAN_HUMAN	P17080	gtp-binding nuclear protein ran	not found in datasets
Sw:KAC_HUMAN	P01834	ig kappa chain c region.	not found in datasets
Sw:MA32_HUMAN	Q07021	pre-mrna splicing factor sf2, p32 subunit.	not found in datasets
Sw:S109_HUMAN	P06702	calgranulin b (mrp-14)	different isoform detected
Sw:MLRN_HUMAN	P24844	myosin regulatory light chain 2	not found in datasets
Sw:POR2_HUMAN	P45880	voltage-dependent anion-selective channel protein 2	different isoform detected
Sw:RET1_HUMAN	P09455	retinol-binding protein I	not found in datasets
Sw:RINI_HUMAN	P13489	placental ribonuclease inhibitor	not found in datasets
Sw:TCTP_HUMAN	P13693	translationally controlled tumor protein (p23)	not found in datasets

Table 4: This list represents the differences in detected proteins compared to the list of overexpressed proteins presented by *Lu* and collaborators. The major part of proteins on this table are not identified in the two different samples. For some proteins different isoforms were detected. Only one good spectrum per sample was measured for IQG1_HUMAN (ras gtpase-activating-like protein iqgap1). The spectrum for this protein could not be clustered.

Protein	pMismatch	Annotation
sw:UCR1_HUMAN	-21.5718933	ubiquinol-cytochrome c reductase complex core protein i
sw:K1CI_HUMAN	-25.9786091	keratin, type i cytoskeletal 9 (cytokeratin 9) (k9) (ck 9).
sw:CAPB_HUMAN	-21.4474485	f-actin capping protein beta subunit (capz beta).
sw:LAM1_HUMAN	-19.3150724	lamin b1.
sw:IMMT_HUMAN	-23.511402	mitochondrial inner membrane protein (mitofilin) (p87/89).
sw:ACY1_HUMAN	-21.8442881	aminoacylase-1 (ec 3.5.1.14) (n-acyl-l-amino-acid amidohydrolase) (acy-1).
sw:LEG3_HUMAN	-25.9721977	galectin-3 (galactose-specific lectin 3)
sw:PSA2_HUMAN	-19.7282655	proteasome subunit alpha type 2 (ec 3.4.25.1)
sw:FIBG_HUMAN	-28.5129776	fibrinogen gamma chain precursor (pro2061).
sw:CA26_HUMAN	-19.0450653	collagen alpha 2(vi) chain precursor.

Table 5: This table shows proteins expressed in the carcinoma sample that have not been found in the control. Red marked proteins are not described to be differentially expressed in the work published by *Lu et al*.

As demonstrated on Figures 28-30, it is almost impossible to match gels from different samples accurately using traditional gel matching software. Almost 40% of the proteins were assigned to be overexpressed by mistake. Gels subject of gel comparisons have to be very similar to obtain good results. Often, this is not the case. Better and more reliable results are obtained by comparing differences in identifications of proteins. Here, all proteins found exclusively in one or the other sample have been assigned correctly. However, quantitative and reliable results are obtained when comparing differences in protein expression on the level of spectral information. Relating spot volumes to protein identifications is not possible when comparing protein identifications due to the non-optimal identification rate of peptide mass fingerprint search algorithms.

The computationally expensive part of the whole evaluation was the clustering of spectra of the datasets, which was conveniently done overnight without operator involvement. The gel matching on the other hand took around 2 minutes to complete. Thus, our method outperforms traditional gel comparison software in terms of speed and accuracy. Gel image comparison algorithms work fine for 60% of the spots and their proteins. However, the amount of false positive annotations is too large to derive robust conclusions.

The disadvantage of the alternative method – having few positive hits when comparing counts of protein identifications – is circumvented by comparing directly the spectral information as it is done using our method.

6. Conclusions

We established a new algorithm to accurately measure the similarity between MALDI-TOF measured mass spectrometric data. Our algorithm is a combination of two orthogonal correlation measures, a correlation for monoisotopic peak masses and a correlation for relative levels over noise of the peaks. The superior performance of our algorithm is shown in chapter 5.1.

We used this algorithm to compare large data sets of mass spectrometric data. As it is described in chapter 5.2 we used a graph theoretical clustering algorithm to robustly group similar spectra together. We established a method to exclude noise peaks from sets of mass spectrometric data. Peaks with known masses are filtered from the spectra after annotation. Matrix derived peaks and chemical noise are excluded using a method implemented in connection with the clustering procedure that analyses spot locations derived from a spectra cluster on gels and decides whether the set of locations drawn from a gel represents a random sampling of gel spots. This newly established clustering procedure shows excellent performance and filters most of the noise out of the spectra set.

We were able to extract essential peak information from clusters of spectra and to generate consensus spectra. We build a database of consensus spectra derived from clustering of more than 100'000 spectra from four different organisms. This database contains ~1700 consensus spectra that account for the identifications of 1250 different proteins. This peptide/spectra database served for two purposes: we could extract knowledge from matched peptides and compare it to unmatched peptides, and we could compare spectra from any dataset to this reference library of consensus spectra and identify the underlying protein by pair–wise spectra comparison.

The combined knowledge of peptide analysis drawn from the library was implemented in the peptide mass fingerprinting algorithm we employ for identification purposes. We are able to search the databases and identify proteins in spectra with an accuracy of 99.75% true hits at the without loss of sensitivity.

However, it was not possible to extract sufficient knowledge from the comparison of unmatched peptides versus matched peptides to allow prediction of ionization characteristics.

For every noise free cluster, a region on the gel can be defined that represents one (or few) protein(s). Missing spot information is added to already defined regions following two concepts. Spots are added due to its immediate proximity to an existing cluster, or spots are added to a region by having the same surrounding region matched on other gels. After complete matching of gels, cluster regions cover up to 90% of the spots of the gel, depending on the quality of the gel and the spectra. Our new gel matching method is completely independent from protein identifications and/or gel image matching. Accuracy of the gel matching using spectral information is increased in comparison to the traditional methods of gel matching.

As we have now a tool for the complete analysis of gels, spots and its spectral content, we are able to compare large data sets of proteomics experiments. Data sets of different samples can be compared in order to identify differences in protein expression. Therefore, the proposed set of methods represents a new toolbox for the complete analysis of proteomics data, which is fully automated, parameter-free and robust.

An example of the usage of the method is given in chapter 5.9.1.

7. References

1. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **19**, 1720-30 (1999).
2. Anderson, L. & Seilhamer, J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18**, 533-7 (1997).
3. Lockhart, D. J. *et al*. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**, 1675-80 (1996).
4. Gershon, D. Proteomics technologies: probing the proteome. *Nature* **424**, 581-7 (2003).
5. Kenrick, K. G. & Margolis, J. Isoelectric focusing and gradient gel electrophoresis: a two-dimensional technique. *Anal Biochem* **33**, 204-7 (1970).
6. Klose, J. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26**, 231-43 (1975).
7. O'Farrell, P. H. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250**, 4007-21 (1975).
8. Scheele, G. A. Two-dimensional gel analysis of soluble proteins. Charaterization of guinea pig exocrine pancreatic proteins. *J Biol Chem* **250**, 5375-85 (1975).
9. Bjellqvist, B. *et al*. Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J Biochem Biophys Methods* **6**, 317-39 (1982).
10. Cleveland, D. W., Fischer, S. G., Kirschner, M. W. & Laemmli, U. K. Peptide mapping by limited proteolysis in sodium dodecyl sulfate and analysis by gel electrophoresis. *J Biol Chem* **252**, 1102-6 (1977).
11. Yates, J. R., 3rd, Speicher, S., Griffin, P. R. & Hunkapiller, T. Peptide mass maps: a highly informative approach to protein identification. *Anal Biochem* **214**, 397-408 (1993).
12. Pappin, D. J., Hojrup, P. & Bleasby, A. J. Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* **3**, 327-32 (1993).
13. Mann, M., Hojrup, P. & Roepstorff, P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom* **22**, 338-45 (1993).
14. Henzel, W. J. *et al*. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc Natl Acad Sci U S A* **90**, 5011-5 (1993).
15. James, P., Quadroni, M., Carafoli, E. & Gonnet, G. Protein identification by mass profile fingerprinting. *Biochem Biophys Res Commun* **195**, 58-64 (1993).
16. Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **60**, 2299-301 (1988).
17. Karas, M., Bachmann, D., Bahr, U. & Hillenkamp, F. Matrix-assisted ultraviolet laser desorption of non volatile compounds. *Int J Mass Spectrom Ion Processes* **78**, 53-68 (1987).
18. Spengler, B. in *Proteome Research: Mass Spectrometry* (ed. James, P.) 39-40 (Springer-Verlag, Berlin Heidelberg, 2001).

19. Brown, R. S. & Lennon, J. J. Mass resolution improvement by incorporation of pulsed ion extraction in a matrix-assisted laser desorption/ionization linear time-of-flight mass spectrometer. *Anal Chem* **67**, 1998-2003 (1995).
20. Whittal, R. M. & Li, L. High-resolution matrix-assisted laser desorption/ionization in a linear time-of-flight mass spectrometer. *Anal Chem* **67**, 1950-4 (1995).
21. Wiley, W. C. & McLaren, I. H. Time-of-flight mass spectrometer with improved resolution. *Rev of Sci Instrument* **26**, 1150-1157 (1955).
22. Appel, R. D. *et al*. The MELANIE project: from a biopsy to automatic protein map interpretation by computer. *Electrophoresis* **12**, 722-35 (1991).
23. Cordwell, S. J. *et al*. Cross-species identification of proteins separated by two-dimensional gel electrophoresis using matrix-assisted laser desorption ionisation/time-of-flight mass spectrometry and amino acid composition. *Electrophoresis* **16**, 438-43 (1995).
24. Langen, H. *et al*. From genome to proteome: protein map of Haemophilus influenzae. *Electrophoresis* **18**, 1184-92 (1997).
25. Fountoulakis, M., Langen, H., Evers, S., Gray, C. & Takacs, B. Two-dimensional map of Haemophilus influenzae following protein enrichment by heparin chromatography. *Electrophoresis* **18**, 1193-202 (1997).
26. Schwartz, S. A., Weil, R. J., Johnson, M. D., Toms, S. A. & Caprioli, R. M. Protein profiling in brain tumors using mass spectrometry: feasibility of a new technique for the analysis of protein expression. *Clin Cancer Res* **10**, 981-7 (2004).
27. Chaurand, P. *et al*. Integrating histology and imaging mass spectrometry. *Anal Chem* **76**, 1145-55 (2004).
28. Gooley, A. A. & Packer, N. H. in *Proteome Reasearch: New Frontiers in Functional Genomics* (eds. Wilkins, M. R., Williams, K. L., Appel, R. D. & Hochstrasser, D. F.) 70-76 (Springer-Verlag, Berlin Heidelberg, 1997).
29. Quadroni, M. in *Proteome Research: Mass Spectrometry* (ed. James, P.) 190-197 (Springer-Verlag, Berlin Heidelberg, 2001).
30. Fountoulakis, M. & Langen, H. Identification of proteins by matrix-assisted laser desorption ionization-mass spectrometry following in-gel digestion in low-salt, nonvolatile buffer and simplified peptide recovery. *Anal Biochem* **250**, 153-6 (1997).
31. Apweiler, R. *et al*. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32 Database issue**, D115-9 (2004).
32. Press, W. H., Teukolsky, S. A., Flannery, B. P. & Vetterling, W. T. in *Numerical Recipes in C* 681-688 (Cambridge University Press, Cambridge, 1992).
33. Berndt, P., Hobohm, U. & Langen, H. Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. *Electrophoresis* **20**, 3521-6 (1999).
34. Weisstein, E. W. in *Gaussian Function* (From MathWorld--A Wolfram Web Resource, 1999).
35. Press, W. H., Teukolsky, S. A., Flannery, B. P. & Vetterling, W. T. in *Numerical Recipes in C* 545-546 (Cambridge University Press, Cambridge, 1992).
36. Alfassi, Z. B. On the normalization of a mass spectrum for comparison of two spectra. *J Am Soc Mass Spectrom* **15**, 385-387 (2004).

37. Press, W. H., Teukolsky, S. A., Flannery, B. P. & Vetterling, W. T. in *Numerical Recipes in C* 639-642 (Cambridge University Press, Cambridge, 1992).
38. Sokal, R. R. & Michener, C. D. A statistical method for evaluating systematic relationships. *University of Kansas scientific Bulletin* **28**, 1409-1438 (1958).
39. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. in *Biological Sequence Analysis* 166-168 (Cambridge University Press, Cambridge, 1998).
40. Knuth, D. E. in *The Art of Computer Programming; Fundamental Algorithms* 363-372 (Addison-Wesley, 1997).
41. Monigatti, F. & Berndt, P. Algorithm for accurate similarity measurements of peptide mass fingerprints and its application. *J Am Soc Mass Spectrom*, in Press (2004).
42. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105-32 (1982).
43. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. in *Biological sequence analysis* 308-309 (Cambridge University Press, Cambridge, 1998).
44. Shannon, C. E. The mathematical theory of communication. 1963. *MD Comput* **14**, 306-17 (1997).
45. Press, W. H., Teukolsky, S. A., Flannery, B. P. & Vetterling, W. T. in *Numerical Recipes in C* 620-626 (Cambridge University Press, Cambridge, 1992).
46. Weisstein, E. W. in *Voronoi Diagram* (From MathWorld--A Wolfram Web Resource, 1999).
47. Weisstein, E. W. in *K-Means Clustering Algorithm* (From MathWorld--A Wolfram Web Resource, 1999).
48. Beer, I., Barnea, E., Ziv, T. & Admon, A. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* **4**, 950-60 (2004).
49. Stein, S. E. & Scott, D. R. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification. *J Am Soc Mass Spectrom* **5**, 859-866 (1994).
50. LEDA: Library for Efficient Data Types and Algorithms. *Algorithmic Solutions Software GmbH* (2004).
51. Thiede, B. *et al*. Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. *Rapid Commun Mass Spectrom* **14**, 496-502 (2000).
52. Weisstein, E. W. in *Binomial Coefficient* (From MathWorld--A Wolfram Web Resource, 1999).
53. Harrison, A. G. The gas-phase basicities and proton affinities of amino acids and peptides. *Mass Spectrometry Reviews* **16**, 201-217 (1997).
54. Pashkova, A., Moskovets, E. & Karger, B. L. Coumarin Tags for Improved Analysis of Peptides by MALDI-TOF MS and MS/MS. 1. Enhancement in MALDI MS Signal Intensities. *Anal Chem* **76**, 4550-4557 (2004).
55. Valero, M., Giralt, E. & Andreu, D. An investigation of residue-specific contributions to peptide desorption in MALDI-TOF mass spectrometry. *Lett. Pep. Sci.* **6**, 109-115 (1999).
56. Cohen, S. L. & Chait, B. T. Influence of Matrix Solution Conditions on the MALDI-MS Analysis of Peptides and Proteins. *Anal Chem* **68**, 31-37 (1996).

57. Olumee, Z., Sadeghi, M., Tang, X. D. & A., V. Amino acid composition and wavelength effects in matrix-assisted laser desorption/ionization. *Rapid Commun Mass Spectrom* **9**, 744-752 (1995).
58. Lu, Z., Hu, L., Evers, S., Chen, J. & Shen, Y. Differential Expression Profiling of Human Pancreatic Adenocarcinoma and Healthy Pancreatic Tissue. (2004).

8. Acknowledgements

First of all, I want to thank Hanno Langen for giving me the opportunity to perform my Ph.D. thesis in his lab. During the three years of my thesis, I very much appreciated the friendly atmosphere and the scientific enthusiasm of which he is one of the main contributors. Despite his very tight schedule, he had always time to discuss the progress of my work.

Numerous people contributed with their advices and helpful suggestions to the progress of my work, but Peter Berndt is the person to whom I want to express my greatest gratefulness and respect. He supervised my thesis and supported me during the three years. Peter truly deserves the title of a genius doctor-father. Not only he is a master programmer, he is also a true biologist, the incarnation of a Bioinformatician so to say. Peter never lost patience with me even though anybody else would have and he is responsible for the huge increase of my knowledge about computers and biology by sharing some of his with me.

I also want to thank Torsten Schwede who supported my thesis at the Biozentrum. He and his group welcomed me as if belonged to them, even though I was an outsider. Torsten gave me the opportunity to profit from his immense knowledge about protein structures that contributed to my understanding of biological processes and was the basis for some fruitful discussions about ionization processes.

During the three years of my thesis, Nikos Berntenis shared the lab with me and I want to express my gratitude to him for being a good friend. We spent many hours of discussing algorithms, politics, differences between Greece and Switzerland and advantages/disadvantages of using certain not mentioned software. A special thank also goes to Helene Meistermann. I tested my software on some of the numerous gels she made (and it worked!). Additionally, I want to thank all members of the Roche Basel proteomics laboratory, who helped me during my thesis, scientifically and non-scientifically.

Especially my parents and my friends supported me and therefore also my thesis in the outside of the laboratory and I want to express my gratitude to them. Finally yet

importantly, I want to thank Sabine who stands by me in good times as well as in not so good ones.

Appendix

Derivation of the spectra similarity algorithm

$$F(t) = \int_{-\infty}^{+\infty} f(x) e^{+itx} dx \quad (1)$$

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} a e^{-\frac{(x-m_0)^2}{2s^2}} e^{+itx} dx \quad (2)$$

Formulation (2) shows the Fourier transform of a Gaussian.

$$F(t) = \frac{1}{\sqrt{2\pi}} a e^{-\frac{m_0^2}{2s^2}} \int_{-\infty}^{+\infty} e^{\frac{-x^2+m_0 x}{2s^2} + \frac{2its2x}{2s2}} dx \quad (3) \qquad c = 2(m_0 + its^2)$$

$$F(t) = \frac{1}{\sqrt{2\pi}} a e^{-\frac{m_0^2}{2s^2}} \int_{-\infty}^{+\infty} e^{\frac{-x^2+cx}{2s^2}} dx \quad (4)$$

$$F(t) = \frac{1}{\sqrt{2\pi}} a e^{-\frac{4m_0^2+c^2}{8s^2}} \int_{-\infty}^{+\infty} e^{\frac{-(x-\frac{c}{2})^2}{2s^2}} dx \quad (5)$$

$$u = (x - \frac{c}{2}) \qquad \frac{du}{dx} = 1 \qquad du = dx$$

$$F(t) = \frac{1}{\sqrt{2\pi}} a e^{-\frac{4m_0^2+c^2}{8s^2}} \int_{-\infty}^{+\infty} e^{\frac{-u^2}{2s^2}} dx \quad (6)$$

The integral $\int_0^{\infty} e^{-au^2}$ is defined as $\frac{1}{2}\sqrt{\frac{\pi}{a}}$ for $a > 0$. As the integration boarder range from $-\infty$ to $+\infty$ the term is multiplied by the factor 2, which leads to the simplified form shown below.

$$F(t) = sae^{-\frac{s^2t^2}{2} + im_0 t} \quad (7)$$

Formulation (7) shows the final form of a Fourier transformed Gaussian *F(t)*, which again yields a Gaussian.

To calculate the correlation between all the peaks from the spectrum the sum of the Fourier-Transformation of each peak has to be formed:

$$\sum F(t,i) = \sum_{i=1}^{i=N} s_i a_i e^{-\frac{s_i^2 t^2}{2} + i m_0 t} \quad (8)$$

Correlations for each peak with all the other peaks are calculated by multiplying the Fourier sums of the first dimension with the Fourier sums of the second dimension. This product yields

$$\sum\sum F(t,i,j) = \left[\sum_{i=1}^{i=N_i} s_i a_i e^{-\frac{s_i^2 t^2}{2} + i m_0 t}\right]\left[\sum_{j=1}^{j=N_j} s_j a_j e^{-\frac{s_j^2 t^2}{2} + i m_0 t}\right] \quad (9)$$

and looks like that in a simplified form:

$$\sum_{i=1}^{i=N_i}\sum_{j=1}^{j=N_j} s_i s_j a_i a_j e^{\frac{-t^2(s_i^2+s_j^2)}{2} + i(m_{0i} - m_{0j})t} \quad (10)$$

In fact, the formulation shown above is the product of the two conjugates. To obtain the correlation function, the product has to be re-transformed out of Fourier space into normal space. This procedure is explained in detail in the following calculations (11-17).

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \sum\sum F(t,i,j) e^{-ixt} dt \quad (11)$$

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \sum_{i=1}^{i=N_i}\sum_{j=1}^{j=N_j} s_i s_j a_i a_j e^{\frac{-t^2(s_i^2+s_j^2)}{2} + i(m_{0i} - m_{0j})t} e^{-ixt} dt \quad (12)$$

$$A = \sum_{j=1}^{i=N_i}\sum_{j=1}^{j=N_j} s_i s_j a_i a_j \qquad a = s_i^2 + s_j^2 \qquad b = 2i(m_{0i} - m_{0j} - x)$$

$$f(x) = \frac{1}{\sqrt{2\pi}} A e^{\frac{b^2}{8a}} \int_{-\infty}^{+\infty} e^{-\left(\frac{at}{\sqrt{a}} - \frac{b}{2\sqrt{a}}\right)^2} dt \quad (13)$$

$$u = \frac{at}{\sqrt{a}} - \frac{b}{2\sqrt{a}} \qquad \frac{du}{dt} = \frac{a}{\sqrt{a}} \qquad dt = \frac{du\sqrt{a}}{a}$$

$$f(x) = \frac{2}{\sqrt{2\pi}} A \frac{1}{\sqrt{a}} e^{\frac{b^2}{8a}} \int_{0}^{+\infty} e^{-\frac{u^2}{2}} \quad (14)$$

$$f(x) = \frac{1}{2}\sqrt{2\pi}\,\frac{2}{\sqrt{2\pi}} A \frac{1}{\sqrt{a}} e^{\frac{b^2}{8a}} \quad (15) \qquad f(x) = A \frac{1}{\sqrt{a}} e^{\frac{b^2}{8a}} \quad (16)$$

$$f(x) = \sum_{i=1}^{i=N_i} \sum_{j=1}^{j=N_j} a_i a_j s_i s_j \frac{1}{\sqrt{s_i^2 + s_j^2}} e^{\frac{-(m_{0i} - m_{0j} - x)^2}{2(s_i^2 + s_j^2)}} \quad (17)$$

Curriculum vitae

Personal Information	Nationality: Swiss Date of Birth: February 4th, 1977 Languages: German, English, French
Research Experience	*2001 – 2004* **Ph. D. thesis** Roche Center for Medical Genomics, Basel Performed in the proteomics laboratory of PD Dr. Hanno Langen under supervision of Dr. Peter Berndt and Prof. Dr. Torsten Schwede Topic: "Algorithms for the analysis of MALDI peptide mass fingerprint spectra for proteomics" *2000 – 2001* **Master thesis** Swiss Institute of Bioinformatics, Geneva Performed in the SWISS-Prot group of Amos Bairoch Topic: "Prediction of tyrosine sulfation in protein sequences using Hidden Markov Models"
Education	*1997 – 2001* Studies in Biology II (molecular biology) at the Biozentrum of the University of Basel *2001* Diploma in Biology II, specialization in biochemistry *10/1999* Second antediploma in Organic Chemistry, Physical Chemistry and Biology *10/1998* First antediploma in Mathematics, Physics and Inorganic Chemistry *1996* Maturity (type B) at the Kantonsschule Chur