Chapter 4

Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants

4.1 Introduction

Proteomics *inter-alia* focuses on the identification of peptides/proteins in complex biological samples (7). Before the identification of the complex constituents, several separation steps are required to reduce the sample complexity. The classical separation method is the two-dimensional gel electrophoresis (11; 12; 13; 14), followed by excision of the detected spots from the gel, digestion with sequence specific proteases and extraction of the cleaved proteins (15; 16). Mass Spectrometric (MS) analysis (6; 8; 9; 10; 114; 115) of the resulting mixture of peptides yields a *peptide mass fingerprint* (PMF): a set of measured molecular masses of the proteolytic peptides derived from the analysed protein (17; 18; 19).

PMF commonly requires matrix assisted laser desorption/ionisation (MALDI) time of flight (TOF) instruments, capable of high throughput analysis of complex samples with minimal pre-cleanup, high femtomolar range sensitivity and accuracy of peptide molecular mass determination up to $5-10$ parts per million (*ppm*) (116; 117; 118; 119). Due to the high ion transmission of the

TOF mass analyzer, this technique is more sensitive compared with other MS techniques. In relation to Electrospray ionisation (ESI) MS (73), MALDI-MS is more tolerant to sample contamination resulting from salts and detergents often present in protein samples due to the separation method. MALDI-MS and ESI-MS have become the standard high throughput proteome analysis techniques in many research laboratories.

The experimental peptide mass lists are generated by the analysis of TOF spectra (58). Ideally, the TOF is proportional to the square root of mass over charge ($\sqrt{m/z}$). Thus, in order to transform the spectrum from TOF into m/z, two calibration constants A and B are necessary. These can be derived by measuring the flight times t of at least two different ions with known masses and fitting them such that $TOF \approx A\sqrt{\frac{m}{z}} + B$. After the transformation from time into m/z, the mono-isotopic peptide signals in the spectrum are identified and their intensity is determined by computational methods (30; 35; 82; 112). The lists of the first mono-isotopic peptide peaks – further called *peak-lists* – are used to identify the protein of interest. In order to assign the PMF to a protein in a sequence database, database search algorithms use the match (within a given measurement accuracy) of theoretical peptide masses computed from protein sequence databases (104) with observed MS masses (18; 19).

Usually the scoring schemes model the mass frequencies of the proteins and peptides in the sequence databases (27; 29; 30; 31). Other properties to be considered include the different sensitivity of detection for individual peptides, known protein modifications, and/or possible mutations (33; 34; 35; 39), although generally, all popular search scores depend on the *precise* assignment of experimental to theoretical peptide masses.

4.1.1   Two novel calibration methods

In a high throughput setting (56; 57), where the samples are placed on a moving sample support, the calibration coefficients for transforming the TOF into m/z differ depending on sample position. This is due to deviations in plate flatness, sample topography changing the size of the acceleration region (40; 56), and alterations in the strength of the electric field on the sample support borders which

influences the drift velocity of the ions (58). Thus, when calibration constants determined from one position on the sample support are used to calibrate TOF spectra acquired on other positions (a procedure known as *external calibration*), the determined m/z values have errors of up to 500 ppm.

Calibration is usually performed using external (40; 120; 121) or internal calibrants (42; 43), which rely on known masses to calibrate the spectra to common co-ordinates. It must be stressed, that in some cases the signal of a reference compounds might be suppressed by the analyte molecules, thus precluding internal calibration. In other cases, the reference signal may partially overlap with an analyte signal, resulting in an erroneous assignment. A third category of calibration methods is based on the peptide mass rule (30; 35). A major advantage of the latter method is that no internal calibrants are required to calibrate the peak-lists. The limitation of this method is it's sensitivity to the presence of non-peptide peaks in the spectra, and that it completely fails if the number of peptide peaks in peak-lists are small (30; 35; 43). Therefore, in practice this method usually is used only to pre-calibrate (30) or to support the results of internal calibration (43; 112).

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

4.1.2 Evaluating the methods

To demonstrate the accuracy of our methods, we studied one sample set of 380 mass spectra, consisting of a part of the *Arabidopsis thaliana* proteome study (106). For this purpose, a MALDI MS sample support in pre-structured (57)

(384-well) microtitre plate format was used. The measurements were performed using the *Autoflex* MALDI-TOF MS (84) instrument.

To compare the performance of calibration methods described here with those already published (43; 112), we used two different data sets. The first set consisted of 1193 spectra deposited on four pre-structured sample supports and measured on a Reflex MALDI-TOF MS (84) instrument (Reflex data set). Spectra were generated via mass spectrometric analysis of the *Rhodopirellula baltica* proteome (unpublished data). The second set was generated in connection with a proteome study of *Mus musculus* and consisted of 1882 spectra deposited on five pre-structured sample supports and measured on an Ultraflex MALDI-TOF MS (84) instrument (Ultraflex data set).

During MS sample preparation of the Ultraflex data set, standard peptides of known masses (human Angiotensin I - $1,296.6853Da$, human ACTH (18-39) $2,465.1989Da$) were added before the measurement to the MS matrix. This was done because the data sets were optimised for the calibration methods, which required the internal calibrants. We examined if the standard peaks could be observed in *more than* 33% of spectra and if so, we removed the peaks matching these masses from the data set. This procedure was applied in order to simulate a data set not optimised for internal calibration.

The *Rhodopirellula* peptide peak-lists were searched against a Pirelulla database (108) with $13,331$ predicted Open Reading Frames (ORFs). The *Mus musculus* samples underwent searches against the *Mus musculus* entries ($69,343$ -sequences) of the NCBI non-redundant protein database (107).

4.2 Methods

4.2.1 Data sets

In this study, we used three data sets generated in different proteome analyses:

1. A bacterial proteome *Rhodopirellula baltica* (unpublished data) ($1,193$ spectra) measured on a Reflex III (84) MALDI-TOF instrument.

2. A mammalian proteome *Mus musclus* $(1,882$ spectra) measured on Ultraflex (84) MALDI-TOF instrument.

3. A plant proteome *Arabidopsis thaliana* (106) measured on an Autoflex (84) MALDI-TOF instrument.

All PMF MS spectra derive from tryptic protein digests of individually excised protein spots. For this purpose, the whole tissue/cell protein extracts of the former mentioned organisms were separated by two-dimensional (2D) gel electrophoresis (13) and visualised with MS compatible Coomassie brilliant blue G250 (106). The MALDI-TOF MS analysis was performed using delayed ion extraction and by employing the MALDI AnchorChip ™targets (Bruker Daltonics, Bremen, Germany). Positively charged ions in the range of $700-4,500m/z$ were recorded. Subsequently, the SNAP algorithm of the XTOF spectrum analysis software (Bruker Daltonics, Bremen, Germany) detected the monoisotopic masses of the measured peptides. The sum of the detected monoisotopic masses constitutes the raw peak-list. Before affine mass calibration, mass measurement errors which can be described by higher order polynomials and determined using external calibration (cf. Methods: External Calibration), were removed. Processed peak-lists were then used for the protein database searches with the Mascot search software (Version 1.8.1) (55), employing a mass accuracy of $\pm 0.1Da$. Methionine oxidation was set as a variable and carbamidomethylation of cysteine residues as fixed modification. We allowed only one missed proteolytic cleavage site in the analysis.

4.2.2 Describing the Mass Measurement Error and predicting the correct mass

A mass difference can be described either in absolute $\Delta_A = m_y - m_x[\mathtt{m/z}]$ or in relative $\Delta_R = (m_y - m_x) \cdot 10^6/m_y[\mathtt{ppm}]$ units. The masses in two peak-lists X, Y were compared to each other and we considered two peaks to *match*, in the case of the absolute error if $\Delta_A < a[m/z]$ and in the case of the relative errors if $\Delta_R < a[ppm]$. If we plotted Δ_A or Δ_R as a function of m_{theo}, we observed, besides a white noise component $\epsilon \propto N(0, \sigma^2)$, a systematic dependence. This

dependence was modelled using a function $\hat{f}(m)$. Given $\hat{f}(m)$ we corrected the experimental masses using the equations:

$$m_{\text{corr}} \quad = \quad \frac{m_{\text{exp}}}{1 - \hat{f}_R(m_{\text{exp}}) \cdot 1/10^6} \ , \ \text{or} \tag{4.1}$$

$$m_{\text{corr}} \quad = \quad m_{\text{exp}} + \hat{f}_A(m_{\text{exp}}), \tag{4.2}$$

depending on whether the relative or absolute error was used, to obtain corrected masses m_{corr}.

4.2.3 Affine mass measurement error model

In the first approximation, the mass measurement error can be described by an affine function $\hat{f}_{A/R}(m_i) = c_1 \cdot m_i + c_0$, where m_i is the mass of the matching peaks. The intercept and slope coefficients of this function can be determined using linear regression.

If only one matching peak was found or the mass range enclosed by the matching masses was small (*e.g.* less than $200Da$), as a remedy one can fix:

- the intercept at 0, if absolute difference $\Delta_A[Da]$,

- the slope coefficient at 0, if relative difference $\Delta_R[ppm]$

and determine the slope or intercept respectively from the data.

To correct the experimental masses m_{exp} we used Equation 4.2 for the absolute differences Δ_A of matching peaks and Equation 4.1 in case of relative differences Δ_R.

The difference between theoretical and measured masses is called a mass measurement error, while the alignment of m_{exp} on m_{theo} an *internal calibration* (35; 122; 123).

4.2.4 Determining ubiquitous masses and their filtering

To determine the abundant masses we computed two histograms for each data set. The origin in the first histogram $\hat{f}_h^1$ is $x_0 = \min(M) - h$ and of the second

histogram $\hat{f}_h^2$ is $x_0 = \min{(M)} - h/2$, where M are all masses in the data set and the bandwidth h equals the measurement accuracy (in Da). We divided the range of M into bins of *bandwidth h*

$$B_j = [x_0 + (j-1)h, x_0 + jh], \quad \text{with} \quad j \in 1, \dots, l , \tag{4.3}$$

where $l = (\max(M) - x_0) \bmod h$. Formally the histogram of counts f is given by (124)

$$\hat{f}_h(x) = \sum_{i=1}^{n} \sum_{j}^{l} I(X_i \in B_j) I(x \in B_j) , \tag{4.4}$$

where n represented the number of masses in M. If a bin had more counts than a given threshold, the average mass $\bar{m}$ of all peaks in the bin was computed. In the case of two adjacent or overlapping bins B_1, B_2 with a significant number of counts c, we first computed a weighted average of the bin midpoints using the number of counts in each bin as weight

$$m = \frac{m_1 \cdot c_1 + m_2 \cdot c_2}{c_1 + c_2} , \tag{4.5}$$

where m_1 and m_2 are the bin midpoints. Afterwards, the average mass $\bar{m}$ of all peaks in the range $m \pm h/2$ was computed. All peaks with mass $m \in [\bar{m} \pm h/2]$ were subsequently removed from the data set. Using two overlapping histograms allows the detection of clusters that are scattered over two adjacent bins in one of the histograms. Different ways to determine ubiquitous masses were used and reported by Levender et al. (42) and Kreitler (77).

4.2.5 Standard internal calibration - Alignment to a pre-compiled list of calibration masses

Instead of using a predefined list of calibration masses, we chose the calibration masses adaptively. The calibration list consisted of ubiquitous masses determined for the data set (cf. Determining ubiquitous masses). Some of the peaks in the list of ubiquitous masses could be assigned to tryptic autolysis products. These matches were used to calibrate the abundant masses. The peak-lists in the data set were then aligned to the calibrated list of ubiquitous masses.

4.2.6 Filtering of ubiquitous masses prior to database search

We removed ubiquitous masses that occurred in more than 7.7% of peak-lists (42; 43). Filtering of ubiquitous masses was performed on a calibrated set of peak-lists. As a result, we could use a small bandwidth of $h = 0.2Da$ (Equation 4.3) to determine ubiquitous masses. Next, we checked which of them can be assigned with a significant Probability Based Mascot Score ($PBMS$) to a sequence database entry and subsequently removed these masses from the filtering list. Abundant masses assigned to a database entry usually result from proteins multiply detected on a 2D-gel. The multiple identification is due to different localisation of the protein on the 2D-gel caused by: protein modifications (phosphorylation, glycosylation), different splice variants or by partial protein degradation. Finally, we removed all peaks within the range $\pm 0.1Da$ around the ubiquitous masses.

4.2.7 Thin-plate spline

The thin-plate spline is the two-dimensional analogue to the cubic spline in one dimension (59; 125). Let v_i denote one of the error model coefficients, *e.g.* intercept, at a target location (x_i, y_i). A thin-plate spline $f(x, y)$ is a smooth function which interpolates a surface that is fixed at the landmark points $P_i = (x_i, y_i)$ at a specific height h_i. A thin-plate spline interpolation function can be written as

$$f(x, y) = a_1 + a_x x + a_y y + \sum_{i=0}^{p} w_i U(||(x_i, y_i) - (x, y)||) \, , \qquad (4.6)$$

where $U(r) = r^2 \ln(r)$ is the radial basis function with $r = \sqrt{x^2 + y^2}$. This equation is used to predict an unknown v for location (x, y), and is the unique solution (59; 125) which minimises the equation:

$$I[f(x, y)] = \int \int_{R^2} \left(\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial x^2} \right)^2 \right) dxdy \, . \qquad (4.7)$$

This quantity was called the bending energy of the thin-plate spline function. If noise in the determined coefficients v_i is detected, one may wish to relax the exact interpolation requirement (Equation 4.7). This can be accomplished by multiplying equation 4.7 with a *regularization* parameter λ, a positive scalar, and by adding the residual sum of squares, which gives:

$$H[f(x,y)] = \sum_{i=1}^{n}(v_i - f(x_i - y_i))^2 + \lambda \cdot I[f(x,y)] \;. \tag{4.8}$$

Again, as in case of the cubic smoothing spline with the parameter λ, the degree of smoothing can be determined. In our study, we utilised an implementation of the *Thin-Plate Spline* (TPS) interpolation (126), according to Doug Nychka (127).

4.2.8 Linear Regression and Peptide mass Rule algorithm

Wolski et al. (32) defined the distance measure

$$d_\lambda(m_i, m_j) = \begin{cases} |m_i - m_j| \bmod \lambda_{DB} & \text{if} \quad |m_i - m_j| \bmod \lambda_{DB} < 0.5 \\ -(1 - |m_i - m_j| \bmod \lambda_{DB}) & \text{if} \quad |m_i - m_j| \bmod \lambda_{DB} \geq 0.5 \;, \end{cases} \tag{4.9}$$

which computes given λ_{DB} (the average peptide cluster distance for a sequence database DB against which the search is performed, *e.g.* $\lambda_{DB} = 1.000495$) the deviation of a peptide mass difference $|m_i - m_j|$ from the closest monoisotopic mass predicted by the PM-rule (51). If there was a linear dependence between $|m_i - m_j|$ and $d_\lambda(m_i, m_j)$, then it was caused by the slope of the mass measurement error. If we computed all differences $|m_j - m_i|$ and $d_\lambda(m_i, m_j)$ for peak pairs m_i, m_j with $|m_i, m_j| < 1400$, we could determine the slope coefficient c_1 using linear regression, while fixing the intercept to zero (128). In order to make the prediction robust against *e.g.* non-peptide peaks, we used a robust linear regression (113). We removed the slope by multiplying each mass m_i in the peak-list by $(1 - c_1)$. Next, we identified the intercept, which was the average of the distance $d_\lambda(m_i, 0)$, and corrected for it.

4.2.9 External Calibration

In order to model higher order systematic changes of mass dependent differences Δ of experimental m_{exp} and reference masses m_{theo}, the measurements must be evenly distributed over the whole measurement range (70; 120). To model the dependence $\Delta \propto m$ we used a cubic smoothing spline function (129; 130), given by $\Delta = f(m) + \epsilon_i$, where f is a smooth function, and $\epsilon_i \sim N(0, \sigma^2)$.

In our study, we used an implementation of the smoothing spline function, provided by B.D. Ripley and Martin Mächler (based on Fortran code of T. Hastie and R. Tibshirani) as part of the **R**-*stats* package. Other non-parametric regression methods like local polynomial regression (131) generated similar results for all types of instruments used in this study.

To obtain equidistantly spaced measurements of known masses, *External calibration* was employed. Some sample spots on the sample support are dedicated to calibration only. Calibration samples, of polymer mixtures (40), which yield equidistant peaks were used to precisely estimate the mass-dependent difference function.

4.2.10 Similarity/Quality measures for internal calibration

Peak-lists can be easily aligned if they contain many matching peaks and the masses of these peaks span a wide mass range. The alignment of a peak-list pair (X, Y) fails if no matching peaks are found. We described these properties mathematically by the following similarity measure:

$$S_{X,Y} = \left(\sum_{i=1}^{n} \sum_{j=i+1}^{n} |m_j - m_i|^p \right)^{1/p} , \qquad (4.10)$$

where n represented the number of matches, while m_i and m_j were the masses of matching peaks. This measure computed the sum of all mass differences of the matching peaks. The power p could be used to weight the large differences stronger.

4.2.11 Alignment of a set of peak-list using a Minimum Spanning Tree

To align a whole data-set to a single peak-list and to align the peak-lists with the highest similarity given by Equation 4.10, we computed for all peak-lists pairs a distance matrix D by casting the similarities into dissimilarities. This distance matrix can be represented by a complete, weighted graph G, where the vertices V correspond to peak-lists and the edges are weighted with the pairwise dissimilarity. To connect all vertices in the graph G with edges e of maximal similarity, the *Dijkstra-Prim* algorithm for finding the *Minimum Spanning Tree* (132) was implemented. We present here a modified version of this algorithm (see Figure 4.1). The algorithm was modified with respect to the starting conditions. As a starting vertex s we chose a vertex incident to an edge of smallest distance. In addition to the minimum spanning tree T, the algorithm returns also a list of calibration coefficients C, which align all peak-lists V in the data set to the starting vertex (peak-list) s, and a list with connection weights W.

By traversing the edges in T, we reached each vertex in G, starting at s via edges with the highest possible calibration similarity (smallest distance). This is because we picked $D(uv)$ with the smallest possible distance (Figure 4.1, line 5).

To align peak-list v to the starting peak-list s we needed to determine the coefficients $C(v,s)$ of the difference function $\hat{f}(v,s)$ (Equation 4.2). We could obtain them from the coefficients $C(v,u)$ and $C(u,s)$ of the pairwise difference function $\hat{f}(v,u)$ and $\hat{f}(u,s)$ by:

$$C(v,s) = C(v,u) \circ C(u,s) = \begin{cases} c_1^{vs} &=& c_1^{vu} + c_1^{us} + c_1^{vu}c_1^{us} \\ c_0^{vs} &=& c_0^{us} + c_0^{vu} + c_1^{vu}c_0^{us} \end{cases} , \qquad (4.11)$$

where *e.g.* c_1^{us} denotes the slope coefficient, and c_0^{us} the intercept of the function $\hat{f}(u,s)$.

Input: A graph G with m edges; each edge e has a given length $l(e)$.

Initialise:

    1   Pick a vertex s, which is incident to the edge with smallest distance $D(e)$.

    2   Set $U := s$ and let T be a tree with one vertex, namely s.

    3   Set the calibration coefficients C of s zero, $C(s) := (0,0)$.

    4   Set measure of path weight $W(s) := \infty$.

Grow Tree: While $U \neq V$,

    5   Among all edges uv with $u \in U$ and $v \in V \backslash U$ pick that one with smallest $D(uv)$.

    6   Add uv to T and remove it from G by setting $D(uv) = \infty$.

    7   Add v to U.

    8   Compute $C(v,u)$ where u is used as calibration peak-list. Assign $C(v,s) := C(v,u) \circ C(u,s)$.

    9   Set the measure of path weight $W(v,s) = min(S(uv), W(u,s))$ (S - similarity).

Output:

    10   T – which is a maximum spanning tree.

    11   C – which is the calibration list to align all peak-lists (vertices) to the starting peak-list (vertex) s.

    12   W – which are the weights of the path from $s \rightarrow v \in F$.

    13   S – modified similarity matrix.

Figure 4.1: Modified Dijkstra-Prim minimum spanning tree algorithm. The algorithm starts with vertex s (peak-list) belonging to the peak-list pair with smallest distance (line 1) (the standard algorithm starts with an arbitrary pair). In addition to computing the minimum spanning tree T, the algorithm computes the calibration constants $C(v,s)$ (line 8) and the connection weight $W(u)$ (line 9).

Proof: The masses of the peak-list pairs (v, u), as well as (u, s) can be aligned given the $C(v, u)$ and $C(u, s)$ using the equations

$$
\begin{aligned}
m_u &= m_v + \hat{f}_A(v, u) = m_v + c_1^{vu} \cdot m_v + c_0^{vu} \text{ , and} \\
m_s &= m_u + \hat{f}_A(u, s) = m_u + c_1^{us} \cdot m_u + c_0^{us} \text{ (see Equation 4.2).}
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\Delta_A(v, s) &= m_s - m_v \\
&= m_u + c_1^{us} \cdot m_u + c_0^{us} - m_v \\
&= (m_v + c_1^{vu} \cdot m_v + c_0^{vu}) + c_1^{us} \cdot (m_v + c_1^{vu} \cdot m_v + c_0^{vu}) + c_0^{us} - m_v \\
&= \underbrace{(c_1^{us} + c_1^{vu} + c_1^{vu} c_1^{us})}_{c_1^{vs}} \cdot m_v + \underbrace{c_0^{us} + c_0^{vu} + c_1^{vu} c_0^{us}}_{c_0^{vs}} \text{ .}
\end{aligned}
$$

$C(v, s)$ was computed online using Equation 4.11 while growing the tree (Figure 4.1, line 8). Subsequently, the algorithm returned a list C of calibration constants, where $C(v, s)$ described the calibration coefficients allowing to transform peak-list v into the co-ordinate system of the peak-list of origin s.

In order to gain more confidence in the calibration constants in C, the minimum spanning tree algorithm was iterated n times. For computing the consecutive T_i, C_i, W_i, D_i with $i = 2, \ldots, n$ we applied the dissimilarity matrix D_{i-1} and set as a starting vertex $s_i = s_1$ – the vertex incident to the edge of highest similarity in D_1. The returned T_i, C_i, W_i, D_i differed since we removed in iteration $i - 1$ each visited edge (Figure 4.1, line 6).

The calibration constants $C_i(v, s)$ with $i = 1, .., n$ should ideally be the same. It is known that $C_i(v, s)$ differ due to alignment errors. Therefore, we computed a weighted average of the coefficients of the difference model. As weight of each model $C_i(v, s)$ we utilised the smallest pairwise calibration similarity $W_i(v)$ (Figure 4.1, line 9), on the path from s to v:

$$
C_w(v, s) = \frac{\sum_{i=1}^{n} W_i(v) \cdot C_i(v, s)}{\sum_{i=1}^{n} W_i(v)} \text{ .} \tag{4.12}
$$

We applied the calibration constants in C_w to align all peak-lists to the peak-list s.

4.3 Results and Discussion

4.3.1 Internal calibration using a pre-calibrated list of calibration masses

Internal calibration is a widely used method in mass spectrometry. This method fails however, either if no peaks matching known masses are present or if MS peak assignment is false. A detailed description of the application of internal calibration in a high throughput-MS setting, addressing the two points is given by *e.g.* Chamrad et al. (43), Levander et al. (42) and Samuelson et al. (112). In order to avoid the lack of MS peaks matching the known calibration masses the authors used a pre-compiled list, *e.g.* trypsin autolysis peaks and unidentified, frequently observed masses (133).

Chamrad et al. (43) initiated the calibration procedure with searches for matching masses using a relatively large search window and iterated it with an increased accuracy. In this scheme, a large search window allows false assignments for calibration masses to occur more frequently. If a false assignment occurs in the first iteration, then the determined calibration constants are false and the entire calibration would be wrong. In the next round of calibration, where a search for matching masses is performed with a higher mass accuracy, the calibration would also fail. To prevent this, the authors (43; 112) checked the obtained calibration coefficients against the peptide mass rule (PM-rule) (30; 51) and stopped further calibration attempts where they disagreed substantially.

Levander et al. (42) introduced an adaptive method to eliminate low-sensitivity auto-proteolysis trypsin peaks from the calibration mass list if no high-sensitivity trypsin peaks *e.g.* ($842.5099Da$, $1045.5642Da$, $2211.1046Da$) were found to decrease the chance of false matches. Unfortunately, this method could only be applied for "tryptic" calibration peaks.

Figures 4.2 A & B demonstrate the limitations of a calibration list compiled from ubiquitous masses of the whole data set. One can recognise that out of three abundant masses (in red, Figure 4.2 **A**), only two can be practically used for calibration. Specifically, the first and the third abundant mass in the list of ubiquitous masses (Figure 1 A) match simultaneously two peaks in peak-list 3, 4

and 5 (Figure 4.2 B). Thus, out of five peak-lists only three could be calibrated. The second calibration mass is also of no use, since it is the only calibration mass in the peak-lists 1 and 2 (although these peak-lists do contain other shared masses). This illustrates that the usage of a global calibration list may fail to calibrate a set of peak-lists.
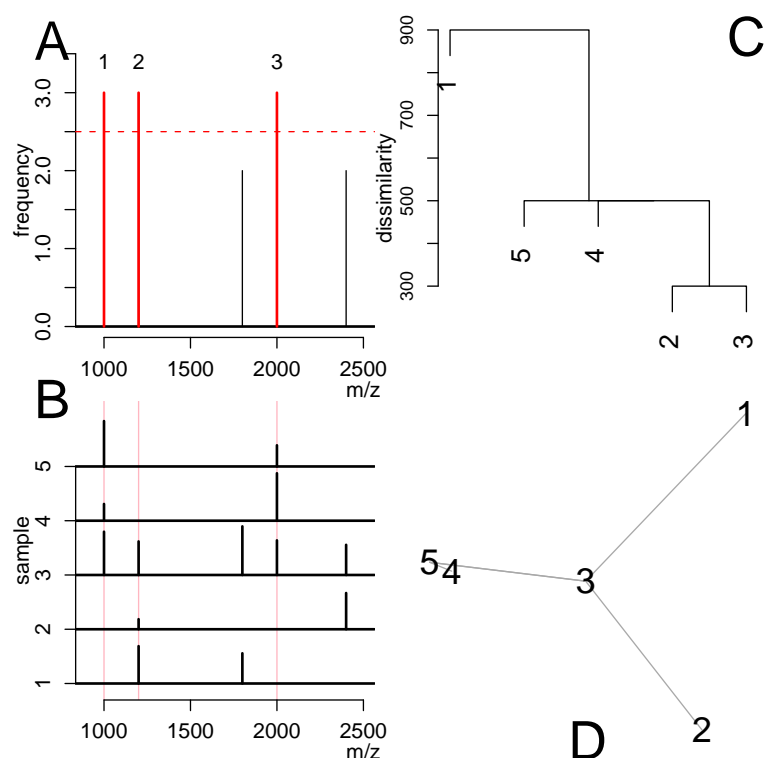

Figure 4.2: **A**: Histogram of masses present in the stick spectra in **B**. In red, marked masses recognised as ubiquitous. **B**: Stick spectra of five hypothetical peak-lists. Red vertical lines mark the position of ubiquitous masses determined using the histogram in **A**. **C**: Single linkage-clustering dendrogram of the peak-lists in **B**. As dissimilarity the mass measurement range (1500 Da) minus the range enclosed by matching peaks was used. **D**: Minimum spanning tree.

It is therefore feasible to address the following questions: How can one obtain a short calibration list to avoid spurious matches while at the same time it matching a sufficient number of peaks in every peak-list of the set? In addition, how can one minimise the initial search window to avoid false matches?

4.3.2 Finding the optimal multiple peak-list alignment using a modified Minimum Spanning Tree algorithm

In order to bypass the limitations imposed by global calibration we used an observation made by Schmidt et al. (53). They noticed that protein samples excised from high-resolution 2D-gels are usually not ideally separated and therefore exhibit local similarities. Compiling a calibration list of abundant masses from a whole data set obtained from a 2D-gel does not differentiate local spectra similarities. For example peak-lists $1, 2$ and 3 (Figure 4.2 **B**) share peaks, which were not recognised as ubiquitous masses and hence not used further for calibration using a global calibration list. The peak-list pairs $(2, 3)$ and $(1, 3)$ shared more than one peak, thus allowing an easy calibration.

We explored the property of local pairwise peak-list similarities for calibration of data sets. To achieve it, we used a modified minimum spanning tree(132) algorithm on the complete weighted graph $G(V, E, d)$, where the vertex set V corresponds to the individual peak-lists and the edges E are weighted by a dissimilarity measure d. We defined the measure between two peak-lists p_1 and p_2 as $d(p_1, p_2) = -s(p_1, p_2)$, where s represented a similarity measure defined in Equation 4.10. This measure not only counts the number of matching peaks, but also weights the mass range enclosed by them. Hence, it also considers that if the matching masses lie very close to each other, the calibration model describes a small mass range only, and can result in a large error when aligning masses that are out of this range.

Using the dissimilarities one can compute a minimum spanning tree (Figure 4.2 D). The algorithm to compute the minimum spanning tree of the peak-list data set starts by choosing a peak-list (named s), which belongs to the peak-list pair of smallest dissimilarity, for example peak-list 2 or 3 in Figure 4.2. This peak-list is the root of the growing tree T (Figure 4.1 line 1). Next, a peak-list v was chosen, which easily could be aligned to peak-list u where u is a part of the growing tree *i.e.* $u \in T$ (Figure 4.1 line 5), for example peak-list $v = 2$ can easily be aligned to peak-list $u = 3$. Using linear regression, we computed the coefficients $c(v, u) = (c_0, c_1)$ of the affine function, modelling the absolute mass differences of the peaks matching in the peak-list pair (v, u). Having these

coefficients one can compute the calibration coefficients $c(v, s)$ using the update rule in Equation 4.11, which described the mass measurement error between the peak-list v and the starting peak-list s. The calibration is not terminated until the whole tree is built. We then added peak-list v to the tree T and have iterated the procedure until all peak-lists were appended to the tree, for example by adding peak-list 4, then 5 and finally 1 to T (Figure 4.2 D).

In the minimum spanning tree algorithm, the vertices are joined by edges of smallest dissimilarity. Consequently, the minimum spanning tree algorithm connects all peak-lists in the data set in the way that the length of the path from the peak-list of origin (root of the tree: peak-list 3 in Figure 4.2 D) to any peak-list in the data set is minimal. The algorithm for computing the agglomerative clustering using the single linkage method (134; 135) works similarly like the minimum spanning tree algorithm and therefore the dendrogram (Figure 4.2 C) provides (as read from bottom to top) the order, by which the peak-list pairs were chosen. The horizontal lines joining two dendrogram tree branches were drawn at the height of the value of the minimal dissimilarity of two peak-lists in either branch.

Finally, the algorithm returns a list of coefficients and a measure of confidence for all peak-lists equalling the smallest similarity in the path from s to v.

Figure 4.3 A demonstrates how the samples on the target are connected by the edges. Green dots (brighter) represent leaves, while blue dots (darker) denote *interior* vertices. The peak-list of origin s is marked with a red cross-hairs (sample position D15). Note that long peak-lists (brighter squares) are *interior* vertices of the minimum spanning tree.

The strip-charts of mass ranges including peaks of the trypsin autolysis products 842.508 and 2, 211.100 are presented in Figure 4.3 C_1 and C_2. One can observe that the minimum spanning tree method works robustly on raw data with a mass measurement error of up to $\pm 0.7 Da$ (black crosses), even if the search for matching peaks when computing the similarities and calibration coefficients was performed within a much smaller window of $\pm 0.45 Da$. Notably, if the maximal error among two peak-lists is much larger than the search window, it is still possible to find a path, thus allowing alignment of two extreme peak-lists.
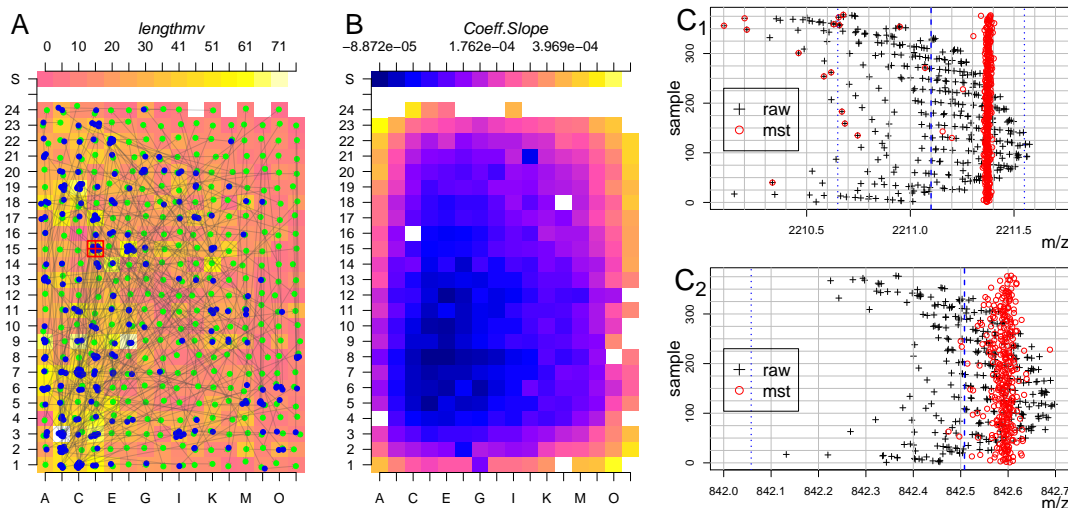
Figure 4.3: **A**: Colour scheme coded peak-list lengths in dependence of the sample support position. Blue dots – *interior* vertex, Green dots – end vertex, white arrows – connecting edges of the minimum spanning tree. The red hair-cross indicates the peak-list of origin s. **B**: Colour scheme coded slope coefficient of the mass- dependent calibration function in relation to sample support position. C_1, C_2: Strip chart of the data set for a mass range of $2210 - 2212Da$ (top) and $842 - 843Da$ (bottom), including the tryptic autolysis peaks 842.508 and $2211.100Da$. Black hair-crosses – masses of peaks before calibration, red circles – masses after calibration. Vertical blue line – the exact position of trypsin autolysis masses 842.508 and $2211.100Da$.

Due to the fact that all peak-lists were aligned to the peak-list of origin s, which did not necessarily match to the theoretical trypsin autolysis masses, a final correction was required to calibrate the whole tree to the theoretical co-ordinate system before database searches (not shown).

4.3.3 Determining the calibration model of the sample support using Thin-Plate Spline interpolation

Because a large part of the mass measurment error is of systematic origin and depends on the sample support position, the mapping of the calibration coefficients across the entire MALDI plate was introduced by Gobom et al. (40) and Moskovets et al. (121). The calibration coefficients were determined using a

standard mixture of peptides with known masses. Subsequently, the calibration coefficients were used during MS analysis in order to correct for the masses measured afterwards on the same plate.

We introduced here a method that derives the calibration model from calibration coefficients acquired from samples, which do not necessarily contain internal standards. Instead of refining the minimum spanning tree calibration model, we chose the peptide mass rule based approach, namely Linear Regression on Peptide Rule (cf. Methods), to obtain the calibration coefficients. The methods based on the peptide mass rule do not rely on the specification of an initial search window or on internal calibrant masses. The peptide rule based calibration method calibrates the peak-lists into the theoretical co-ordinate system and increases the mass accuracy to approximately $0.1Da$, but fails if the peak-list is too short, which indeed could be observed for several samples (Figure 4.4 A and C). Figure 4.4 A provides the color scheme coded slope coefficient c_1 as determined by the peptide rule based calibration method in dependence of the target location. One can observe that some erroneous predictions occur (Figure 4.4 C; black crosses marked by magenta triangles). However, it is unbiased to assume a smooth transition between adjacent positions of the sample support. For example, Figure 4.3 B demonstrates that the slope coefficient of the sample calibration-model obtained by the minimum spanning tree calibration methods increases for samples close to the support border. This change is due to alterations in the electric field E (Equation 4.13) influencing the flight velocity given by

$$u_D = \sqrt{2Es_a\frac{z}{m}}, \tag{4.13}$$

where s_a is the size of the acceleration region, z is the ion charge and m is the mass of the ion. We determined the systematic change of the slope using the thin plate spline interpolation method (59; 127). At first, we computed the thin plate spline with a degree of smoothing $\lambda = 5 \cdot 10^{-2}$ (see Equation 4.8). Calibration models with slope coefficient c_1 that varies more than $\pm 1 \cdot 10^{-4}$ or with intercept coefficient c_0 varying more than $0.2Da$ from the one predicted by the thin plate spline were discarded. Using the remaining calibration models, the thin plate spline was recomputed with smaller degree of smoothing $\lambda = 1 \cdot 10^{-3}$. Figure 4.4
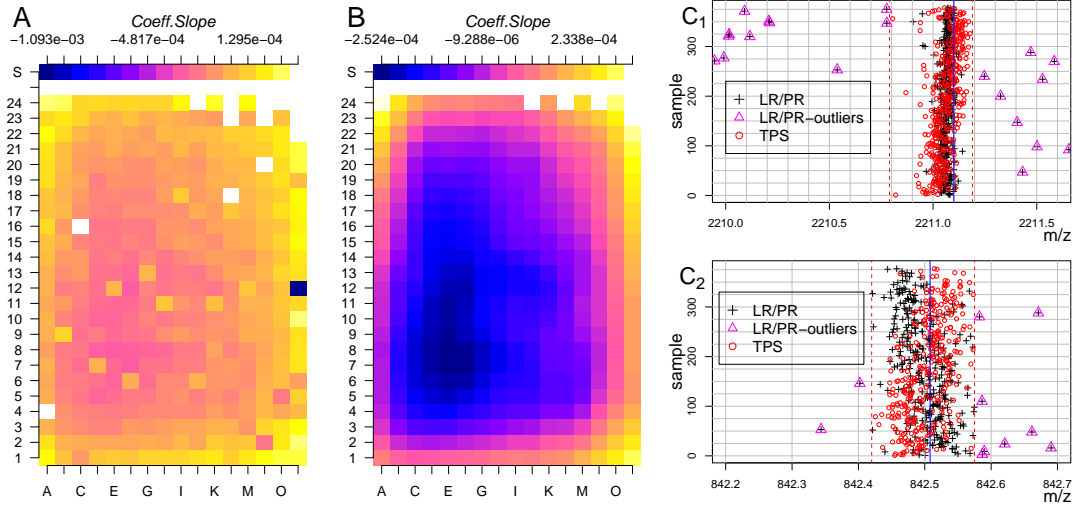
Figure 4.4: **A** : Colour scheme coded slope coefficients $c1$ of the mass measurement error determined by the peptide rule based calibration method. **B**: The slope coefficient as predicted from the refined samples determined by *Thin plate spline with $\lambda = 0.001$.* **C**: Strip chart of the data set for a mass range of $2210 - 2212 Da$ (C_1) and $842 - 843 Da$ (C_2), including the tryptic autolysis peaks 842.508 and $2211.100 Da$. Black crosses – masses of peaks predicted by the peptide rule based calibration method, red circles – masses predicted by the thin plate spline calibration method. Vertical blue line – exact position of trypsin autolysis masses 842.508 and $2211.100 Da$. Dashed red vertical line – mass of the extreme peptide masses after thin plate spline calibration.

B, demonstrates the Colour scheme coded slope coefficient c_1, as estimated by the refined thin plate spline. This model resembles the one generated by the minimum spanning tree method (Figure 4.3 B). We corrected the peak-lists masses (black cross hairs, Figure 4.4 C), using the thin plate spline values as estimates of the slope coefficients, and as intercept estimate we used the average intercept of all coefficients of the refined calibration models to obtain the calibrated masses (red circles).

The thin plate spline method reduced the mass measurement error of a peak-list compared to any other peak-list in the data set (vertical red, dashed line in Figure 4.4 C) down to $0.3 Da$, as compared to $1.5 Da$ for raw data. This is approximately a 5- fold increase of a mass measurement accuracy. This decrease

of the mass measurement error enabled us to utilise the minimum spanning tree algorithm with an accuracy of $\pm 0.15 Da$, reducing further the probability of false assignments of calibration masses. In addition, the histogram of dissimilarities computed for all peak-list pairs (Figure 4.5, A) shows for thin plate spline calibrated data lower values of dissimilarity (in red) as compared to the raw data (in grey), even if the first dissimilarities were computed with a search window of $0.15 Da$ and the second ones with a search window of $0.45 Da$. A subsequent calibration using the minimum spanning tree method decreased further the mass measurement error (Figure 4.5 B).
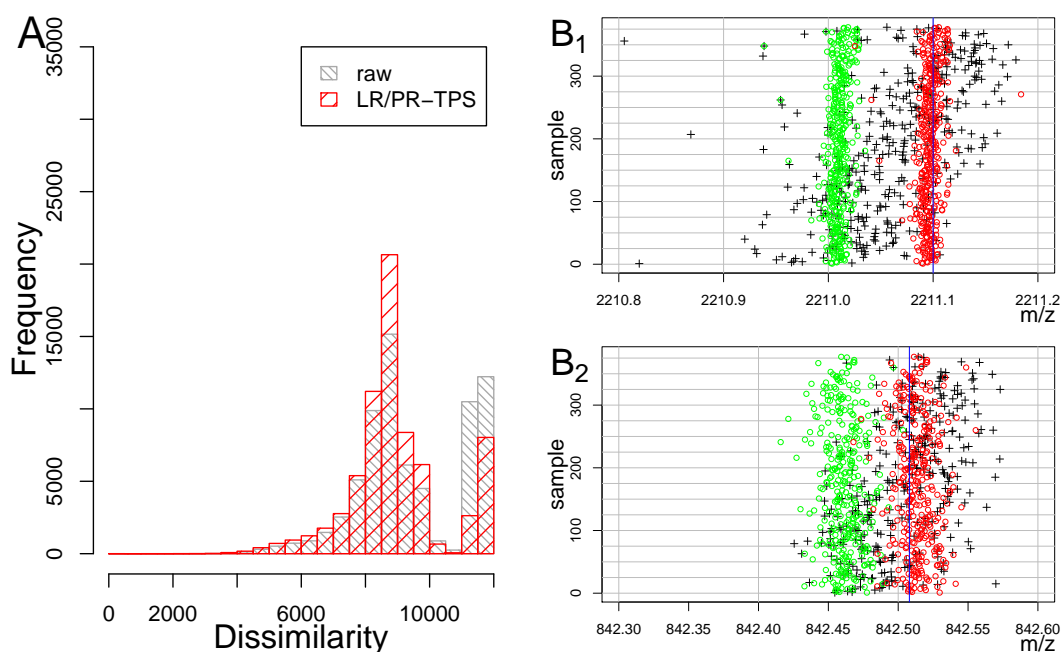

Figure 4.5: **A**: Histogram of pairwise peak-list similarities. In gray – raw data and similarities computed with an accuracy of $\pm 0.4 Da$. In red – similarities computed with accuracy of $\pm 0.15 Da$ using LR/PR and thin plate spline calibrated data. **B**: Strip chart of peak-lists. Grey triangles – masses after thin plate spline calibration, green circles – data after thin plate spline and minimum spanning tree calbration, red circles – data calibrated into the theoretical co-ordinate system, defined by theoretical tryptic autolysis masses (blue vertical lines.)

4.3.4 The mass measurement error

Prior to the calibration, the main error source is due to different drift velocities of the ions causing an increase of the absolute mass measurement error, proportional to mass and best described by the slope coefficient $c_1 \neq 0$ and measured as relative error using parts per million *ppm* (Table 4.1, row 1 and 2). After removal of this error using calibration methods, for example the thin plate spline calibration (Table 4.1, row 3,4) or thin plate spline with subsequent minimum spanning tree calibration (Table 4.1 row 5,6), the main contribution to the mass measurement error was due to peak detection performance. We were aware, however, of systematic changes of the mass measurement error, which can be described using higher order polynomials (120; 122). We have removed higher order terms of the mass measurement error, by applying external calibration before to other calibration procedures (cf. Methods : External Calibration). The change of peak-detection quality was negligible in the range of $500 - 4000 Da$. Figure 4.6, as well as Table 4.1, illustrates that after calibration the absolute mass measurement error was smaller for the peak with higher mass (2211.1) than that of the peak with a lower mass (842.508) if the peak intensity and consequently the Signal to noise ratio remained sufficiently high. Therefore, we performed the database searches by specifying the search window in *Da* instead of *ppm*.

Calibration	Mass	$S_N[Da]$	$S_N[ppm]$
Raw data	842.508	0.1	118
Raw data	2211.1	0.3	135
TPS	842.508	0.03	37
TPS	2211.1	0.057	26
TPS-MST	842.508	0.012	14.5
TPS-MST	2211.1	0.01	4.6

Table 4.1: Mass Measurement Error. Standard deviation (S_N) observed for the trytpic autolysis peaks 842.508 and 2211.1. Raw data; TPS - thin plate spline calibrated data; TPS-MST - The data, which undergone Thin-Plate Spline (TPS)(pre-processing), followed by Maximum Spanning Tree (MST) calibration
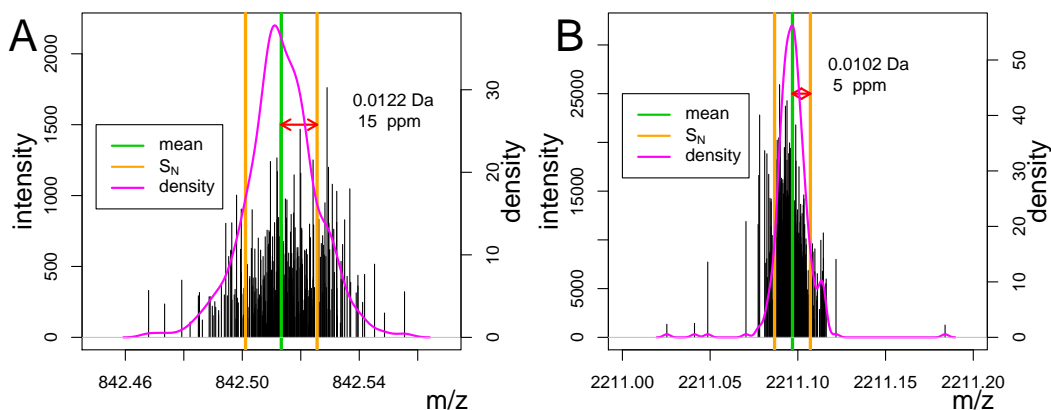
Figure 4.6: Stick spectrum of the merged data set of 380 peak-lists. The black vertical lines represent peaks calibrated using the thin plate spline and minimum spanning tree method. Their height equals their intensity. Green line – average mass of all peaks in the region $842 - 843Da$ (**A**) and $2210.5 - 2211.6Da$ (**B**). The orange vertical lines represent the average mass $\pm$, the standard deviation of the peak masses in each region. Magenta line – density of peak-masses.

4.3.5 The optimal size of the search window

Figure 4.6 and Table 4.1 demonstrate that it is possible to reduce the mass measurement error to approximately $\pm 10ppm$ for most of the peak-lists in a dataset consisting of 380 spectra, by applying the thin plate spline - minimum spanning tree calibration sequence. Nevertheless, in this dataset one can observe peak-lists that do not exhibit such high mass measurement accuracy. Consequently, if the database searches were performed with a search window of $10ppm$, these PLs would not be identified.

The optimal size of the search window was determined by searching of four *internally* calibrated data sets with five different search window sizes, namely $0.5, 0.2, 0.1, 0.05$ and $0.02Da$ using the Mascot (55) search algorithm. The search window of $0.2Da$ generated the highest identification rate. Figure 4.7 shows the relative identification rate (identification rate / max(identification rate) $\cdot 100\%$). Allowing the search window to be larger *e.g.* $0.5Da$, decreases the identification rate by increasing the rate of false negatives, while a smaller window *e.g.* $\pm 0.05Da$ decreases it by rejecting true matches (55). Because the identification rate for

a search window of $0.1Da$ is only slightly worse than one of $0.2Da$, and since it minimizes the risk of false positive matches, we further compared the practical performance of the calibration methods with a search window of $0.1Da$.
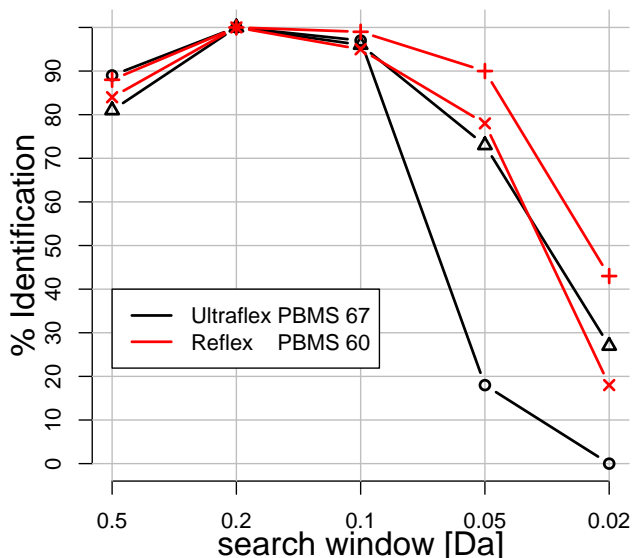

Figure 4.7: The optimal search window. Comparison of the relative identification rates of internally calibrated data (Y-axis) given a search window size of $0.5Da$, $0.2Da$, $0.1Da$, $0.05Da$ and $0.02Da$, respectively (X -axis). Red – Two Reflex (Pirellula) dataset, Black – Two Ultraflex (*Mus Musculus*) datasets

Prior to the database searches we removed all masses that occur in more than 8% of spectra, as it significantly increased the identification rate (42; 43) (cf. Methods – Filtering of ubiquitous masses prior to database search). The sequence data base search was performed using the Mascot (55) search software version 1.8.1. We interfaced the search server from within R using the in-house developed R package `msmascot` (136).

4.3.6   Combining different calibration methods and their comparison

All parameters were fitted to a data set optimised for internal calibration, measured on an *Autoflex* MALDI-TOF MS (84) instrument. We applied the

calibration methods introduced (minimum spanning tree and thin plate spline based calibration) without changing the parameters to two sample sets obtained using two different instruments, namely a Reflex MALDI-TOF MS and a Ultraflex MALDI-TOF MS instrument. This was executed to illustrate that our methods are robust with respect to different instruments even if the parameters were not optimised for the respective machines.

We combined the different pre-calibration and calibration methods resulting in six different calibration sequences (summarised in Table 4.2). We compared the performance of the minimum spanning tree and thin plate spline calibration sequence to the internal calibration (IC), and the peptide rule based calibration methods (LR/PR). Furthermore, we investigated if the identification rate of the thin plate spline based method could be improved further by subsequent internal (TPS-IC) or minimum spanning tree calibration (TPS-MST). The `R` (137) scripts implementing each sequence can be found in the samples directory of the `mscalib` `BioConductor` (138) package.

	Abbreviation	Description
1	LR/PR	peptide rule calibration.
2	IC	internal calibration 450 ppm and 250 ppm.
3	MST	minimum spanning tree calibration.
4	TPS	LR/PR and subsequent thin-plate spline calibration.
5	TPS-IC	thin plate spline calibration and subsequent internal calibration.
6	TPS-MST	thin plate spline calibration and subsequent MST calibration.

Table 4.2: Calibration sequences. LR/PR – linear regression on peptide rule, IC – Internal calibration with two iterations. (Bruker Reflex – mass measurement error window of 450 and 250ppm, Bruker Ultraflex – 250 and 125ppm); MST – minimum spanning tree calibration method computed with an search window of $\pm 0.4 Da$; TPS-IC - Pre-processing (thin plate spline calibration) and subsequent internal calibration with a mass measurement error window of 250ppm; TPS-MST - thin plate spline pre-processing and an minimum spanning tree with a search window of $\pm 0.25 Da$;

The only calibration method for which parameters were optimised with respect to the instrument was the standard internal calibration (IC) method, which

employs a pre-compiled calibration list of theoretical trypsin autolysis peaks and a calibrated set of ubiquitous masses (cf. Methods – Standard internal calibration). In case of the peptide rule based calibration (LR/PR) method we applied an additional filtering of the calibration-models. Only models with an intercept coefficient c_0 satisfying $-0.4Da < c_0 < 0.4Da$ and slope coefficients c_1 with $-5 \cdot 10^{-3} < c_1 < 5 \cdot 10^{-3}$ were kept. In order to avoid falsely calibrated peak-lists we performed the filtering.

The identification rates were defined as the number of identified samples by at least one of the calibration sequences divided by the number of samples submitted for searches

$$\frac{\#\{CS_1 \cup CS_2 \cup \cdots \cup CS_6\}}{\text{number of samples submitted for search}} \, , \qquad (4.14)$$

where CS_i indicates the set of identified samples by one of the calibration sequences (Table 4.2), and $\#\{A\}$ denotes the number of elements in a set A. The identification rates were 74%, 87%, 79%, 85% for the Pirellula (Reflex) data set, with an overall identification rate of 82%, whereas for the *Mus musculus* (Ultraflex) data set they were 51%, 72%, 35%, 51%, 27%, with an overall identification rate of 58%. The lower identification rate of the *Mus musculus* data set can possibly be explained by the fact that it was matched with a larger database. Therefore, more matching peaks are required to make significant assignments to a data base entry.

In order to directly compare the identification rates for both data sets and each calibration sequence, we computed the relative identification rate. It was defined as the ratio of the number of identified samples calibrated by a sequence (numerator) and of the number of identified samples, which could be identified by at least one method (denominator):

$$\frac{\#\{CS_i\}}{\#\{CS_1 \cup CS_2 \cup \cdots \cup CS_6\}} \, . \qquad (4.15)$$

The relative identification rate is indicated by the dots, joined by continuous lines for readability purposes only, in Figure 4.8. The dashed lines denote the average of the sequence coverage of all identified samples. Figure 4.8A presents the results for the four Pirellula data sets, while Figure 4.8B shows the results of five *Mus musculus* data sets.
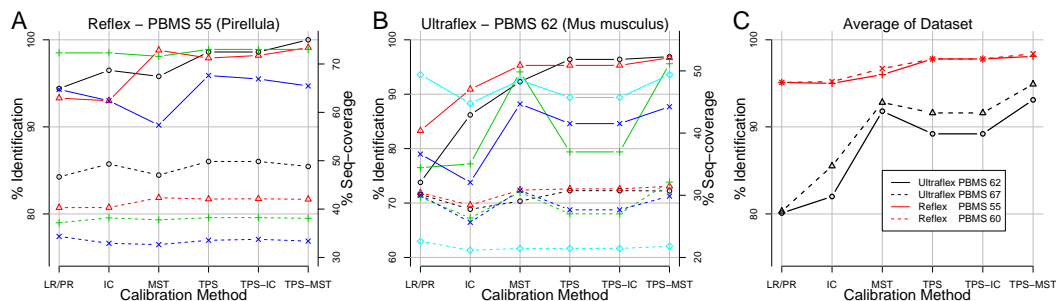
Figure 4.8: Relative identification rate in % (continuous line – left y-axis) and sequence coverage in % (dashed lines - right y-axis). LR/PR – linear regression on peptide rule, IC – two step internal calibration, MST – minimum spanning tree calibration, P – thin plate spline calibration, TPS-IC – thin plate spline calibration and subsequent internal calibration, TPS-MST - thin plate spline calibration and subsequent minimum spanning tree calibration. Panel A - Reflex datasat, Panel B - Ultraflex datset, Panel C - Average of Dataset (see text for details).

Only in one case of one data set was a single calibration sequence: thin plate spline calibration and subsequent minimum spanning tree calibration (TPS-MST) (see Table 4.2) able to identify all peak-lists (100% identification rate) and therefore it completely dominated over the other methods (black line, Figure 4.8 A). In the case of the Ultraflex data set (Figure 4.8 B) we observed that the TPS-MST method had the highest identification rate, while in Reflex data set (Figure 4.8 A) it achieved the highest performance for approximately half of the data sets.

Figure 4.8 C illustrates the averaged relative identification rate of the calibration methods for the Ultraflex and Autoflex data sets. In addition, it demonstrates that the ordering of the calibration methods according to the relative identification rate does not depend on the value of the Probability Based Mowse Score (55) (PBMS) used as identification threshold. The dashed lines (Figure 5) indicate the identification rates obtained for a PBMS 5 units higher than the one used to identify the samples with a 0.5% significance level (continuous lines).

Interestingly, the thin plate spline smoothing method resulted in an overall higher identification rate than the other methods tested on raw data (peptide

rule based calibration, internal calibration, minimum spanning tree-calibration), except for one case of the Ultraflex data set. Furthermore, a combination of the internal calibration with thin plate spline calibration (TPS-IC) did not increase either the sequence coverage (dashed lines) or the identification rate of the thin plate spline method applied alone.

In two out of the four Reflex data sets, the minimum spanning tree method applied on thin plate spline processed data (P-thin plate spline Figure 4.8 A, dashed lines) slightly decreased the sequence coverage indicating a reduction of calibration accuracy. For the Ultraflex data sets, the sequence coverage correlated well with the identification rate and the combination of the thin plate spline with minimum spanning tree method accomplished the highest performance.

Moreover, if similar identification rates of the peptide rule based calibration and the internal calibration were observed, the peptide rule based calibration method provided higher sequence coverage (Figure 4.8 B). This could be explained by the fact that the peptide rule based method calibrated well the peak-lists possessing many peptide peaks. Such peak-lists potentially contain the higher sequence coverage.

4.3.7 The `BioConductor` package `mscalib`

All of the calibration methods are part of the `mscalib` programme, which is available as a *BioConductor* (139) package. The Bioconductor project is an initiative for the collaborative creation of extensible software for computational biology and bioinformatics (138). The scripts carrying out the calibration sequences tested, can be found in the subdirectory */samples* of the package. Furthermore, in the same directory and in the directory */doc* there are two vignettes (140) with detailed descriptions of two selected calibration sequences.

4.4 Conclusions

While the methods described in this study significantly improve the calibration of raw data, they do not perform better than other published calibration routines which reduce the mass measurement error to $10ppm$ or below. The real advantage

of the methods described here is that they are not dependent on the presence of internal or external calibrants, required to correct for the affine component of the mass measurement error. Furthermore, the calibration methods described in this study allow a larger fraction of peak-lists in the datasets to be calibrated than the reference internal calibration method would do.

The thin plate spline method deals with systematic detrimental calibration effects that are due to imperfections in the geometry of the electric field over the MALDI sample plates. Usage of thin plate spline calibration results in up to 10% higher identification rates, at least for the Bruker mass spectrometers, than the internal calibration. The thin plate spline calibration procedure enables, for most of the samples deposited on the sample support, to obtain mass accuracy in the range of $\pm 0.1 Da$. Moreover, the TPS method does not require the presence of internal calibrants since it relies on calibration coefficients acquired from a calibration method based on the peptide mass rule.

The minimum spanning tree method is able to increase the identification rates obtained by the thin plate spline method for protein samples separated by a 2D-Gel electrophoretic procedure. Furthermore, the parameters optimised for one instrument (Autoflex) can be directly utilised for other instruments (Reflex, Ultraflex).

In this work, we have only examined a version of the minimum spanning tree algorithm that builds a single tree for all peak-lists. This is adequate if the data are a set of peak-lists with smooth transitions in the similarity values. If this is not the case, it might be more appropriate to compute a forest of several minimum spanning trees. We have examined, however, only a single peak-list similarity measure (Equation 4.10) for peak-lists calibration. It is possible that better similarity measures can still be generated and subsequently applied for peak-lists calibration.

Complete utilisation of microtitre plates and sample supports is not only rational with respect to increased accuracy of the thin plate spline method, but also with respect to the idea of high throughput experiments – maximal utilisation of energy and resources. Dense excision of spots from 2D-gels not only increases the performance of the minimum spanning tree method, but also identifies novel proteins. Hence, the main contribution of this manuscript is to present two

calibration methods, compatible with the principle of high throughput sample processing and aims to identify a maximum of the proteins resolved on 2D-gels.

However, no single "best-calibration" method exists. Each of the methods utilises different properties of the peak-lists. Consequently, applying these methods in parallel and determining the total (union) of the identified samples provides the highest identification rate.