

Calibration of Mass Spectrometric Peptide Mass Fingerprint Data without specific external or internal calibrants.

Witold E Wolski^{*1/2/5} , Maciej Lalowski³ , Peter Jungblut⁴ , Knut Reinert²

¹Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, D-14195 Berlin, Germany

²Institute for Computer Science, Free University Berlin, Takustr. 9, 14195 Berlin, Germany

³Max Delbrück Center for Molecular Medicine, Robert-Roessle-Str. 10, D-13125 Berlin-Buch, Germany

⁴Max Planck Institute for Infection Biology, Schumannstr. 21-22, D-10117 Berlin

⁵ present address: School of Mathematics and Statistics, Merz Court, University of Newcastle upon Tyne, NE1 7RU, UK

Email: Witold E Wolski^{*} - witek96@users.sourceforge.net; Maciej Lalowski - m.lalowski@mdc-berlin.de; Peter Jungblut - jungblut@mpiib-berlin.mpg.de; Knut Reinert - reinert@inf.fu-berlin.de;

*Corresponding author

Abstract

Background: Peptide Mass Fingerprinting (PMF) is a widely used Mass Spectrometric (MS) method of analysis of proteins and peptides. It relies on the comparison between experimentally determined and theoretical mass spectra. The PMF process requires calibration, usually performed with external or internal calibrants of known molecular masses.

Methods: We have introduced two novel MS calibration methods. The first method utilises the local similarity of peptide maps generated after separation of complex protein samples by 2D-gel electrophoresis. It computes a multiple peak-list alignment of the data set using a modified Minimum Spanning Tree (MST) algorithm. The second method exploits the idea that hundreds of MS samples are measured in parallel on one sample support. It improves the calibration coefficients by applying a two-dimensional Thin Plate Splines (TPS) smoothing algorithm.

Results: We studied the novel calibration methods utilising data generated by three different MALDI-TOF-MS instruments. We demonstrate that a PMF data set can be calibrated without resorting to external or relying on widely occurring internal calibrants. The methods developed here were implemented in R and are part of the BioConductor package `mscalib` available from <http://www.bioconductor.org>.

Conclusions: As compared to other methods, our combined MS spectra calibration strategy increases the peptide/protein identification rate by an additional 5 – 15%.

Background

Proteomics *inter-alia* focuses on the identification [1] of peptides/proteins in complex biological samples. Prior to the identification of the complex constituents, several separation steps are required to reduce the sample complexity. This separation is performed by 2D-gel electrophoresis [2,3], followed by excision of the detected spots from the gel, digestion with sequence specific proteases and extraction of the cleaved proteins [4,5]. Mass Spectrometric (MS) analysis [6–9] 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 [10,11].

PMF commonly requires Matrix Assisted Laser Desorption Ionisation (MALDI) Time of Flight (TOF) instruments, capable of High Throughput (HT) 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*) [12–15]. Due to the high ion transmission of the TOF mass analyser, this technique is more sensitive compared with other MS techniques. In relation to Electrospray ionization (ESI), MALDI is more tolerant to sample contamination, *e.g.* from salts and detergents often present in protein samples due to the separation method. Therefore, MALDI MS has become the standard HT proteome analysis technique in many research laboratories.

The experimental peptide mass lists are obtained from the analysis of TOF spectra [16]. 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 required. 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 [17–20]. The lists of the first mono-isotopic peptide peaks – further called *peak-lists* or short *PLs* – are used to identify the protein of interest. In order to determine the protein sequence, database search algorithms use the match

(within a given measurement accuracy) of theoretical peptide masses computed from protein sequence or transcript databases [21] with observed MS masses [10, 11].

Usually the scoring schemes model the mass frequencies of the proteins and peptides in the sequence databases [18, 22–24]. Other properties to consider include the different sensitivity of detection for individual peptides, known protein modifications, and/or possible mutations [17, 25–27], although generally, all popular search scores depend on the *precise* assignment of experimental to theoretical peptide masses.

Two novel calibration methods

In a HT- setting [28, 29], 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 [28, 30], and alterations in the strength of the electric field on the sample support borders which influences the drift velocity of the ions [16]. 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 [30, 31] or internal calibrants [32, 33], which rely on known masses to calibrate the spectra to common co-ordinates. It must be stressed that when assigning internal reference signals, the signals of a reference compound in some cases 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 [17, 18]. A major advantage of the latter method is that no internal calibrants are required to calibrate the PLs. The limitation of this method is its sensitivity to the presence of non-peptide peaks in the spectra, and that it completely fails if the number of peptide peaks in PLs are small [17, 18, 32]. Therefore, in practice this method usually is used only to pre-calibrate [18] or to support the results of internal calibration [20, 32].

We have developed two novel calibration methods for PMF data. Both calibration methods exploit similarities of PLs due to closeness in the origin of the analysed samples. The first method combines the computation of dissimilarities [34] between PLs with internal calibration. The second method employs spatial statistical methods [35] 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.

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 [36]. For this purpose, a MALDI MS sample support in pre-structured [29] (384-well) microtitre plate format was used. The measurements were performed using the *Autoflex* MALDI-TOF MS [37] instrument.

To compare the performance of calibration methods described here with those already published [20,32], we used two different data sets. The first set consisted of 1193 spectra deposited on 4 pre-structured sample supports and measured on a Reflex MALDI-TOF MS [37] 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 5 pre-structured sample supports and measured on an Ultraflex MALDI-TOF MS [37] instrument (Ultraflex data set).

During MS sample preparation of the Ultraflex data set, standard peptides of known masses (human Angiotensin I- 1296.6853Da, human ACTH (18-39) 2,465.1989Da) were added before the measurement to the MS matrix. This was performed since the data sets were optimized for the calibration methods, which required the internal calibrants. We examined if the standard peaks could be observed in *more than 33%* of spectras 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 optimized for internal calibration.

The *Rhodopirellula* peptide PLs were searched against a Pirelulla database [38] 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 [39].

Results and Discussion

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 HT-MS setting, addressing the two points is given by *e.g.* Chamrad et al. [32], Levander et al. [33] and Samuelson et al. [20]. 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

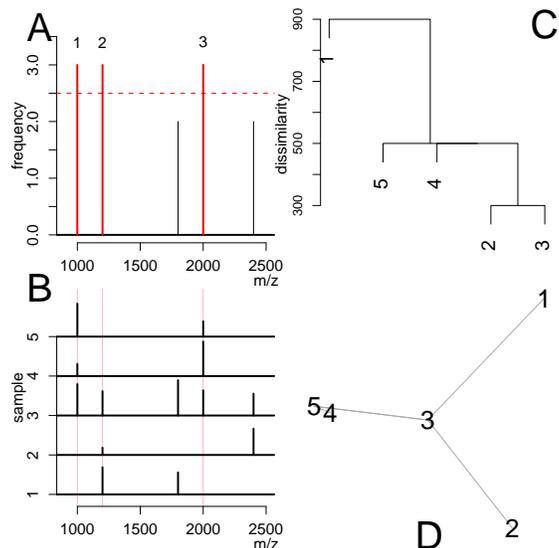


Figure 1: Figure **A**: Histogram of masses present in the stick spectra in Figure **B**. In red, marked masses recognised as ubiquitous. Figure **B**: Stick spectra of five hypothetical PLs. Red vertical lines mark the position of ubiquitous masses determined using the histogram in Figure **A**. Figure **C**: Single linkage-clustering dendrogram of the PLs in Figure **B**. As dissimilarity the mass measurement range (1500 Da) minus the range enclosed by matching peaks. Figure **D**: Minimum spanning tree computed using the dissimilarities.

unidentified, frequently observed masses [40].

Chamrad et al. [32] 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 will also fail. To prevent this, the authors [20,32] checked the obtained calibration coefficients against the peptide mass rule (PM-rule) [18,41] and stopped further calibration attempts in case of disagreement.

Levander et al. [33] 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 can only be applied for “tryptic” calibration peaks.

Figures 1 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 1 **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 PL 3,4 and 5 (Figure 1 B). Thus, out of five PLs, only three could be calibrated. The second calibration mass is also of no use, since it is the only calibration mass in the PLs 1 and 2 (although these PLs do contain other shared masses). This illustrates that the usage of a global calibration list may fail to calibrate a set of PLs.

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

Finding the optimal multiple PL alignment using a modified Minimum Spanning Tree (MST) algorithm

In order to bypass the limitations imposed by global calibration we used an observation made by Schmidt et al. [42]. 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. *e.g.* PLs 1, 2 and 3 (Figure 1 B) share peaks, which were not recognised as ubiquitous masses and hence not used further for calibration using a global calibration list. The PL pairs (2, 3) and (1, 3) shared more than one peak, and could thus easily be calibrated.

We explored the property of local pairwise PL similarities for calibration of data sets. To achieve it, we used a modified *minimum spanning tree* (MST) [43] algorithm on the complete, weighted graph $G(V, E, d)$, where the vertex set V corresponds to the individual PLs, and the edges E are weighted by a dissimilarity measure d . We defined the measure between two PLs 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 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 MST (Figure 1 D). The algorithm to compute the MST of the PL data set starts by choosing a PL (named s), which belongs to the PL pair of smallest dissimilarity, *e.g.* PL 2 or 3 in Figure 1. This PL is the root of the growing tree T (Figure 8 line 1). Next, a PL v was chosen, which easily could be aligned to PL u where u is a part of the growing tree *i.e.* $u \in T$ (Figure 8 line 5), *e.g.* PL $v = 2$ can easily be aligned to PL $u = 3$. 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 PL pair (v, u) . Having these coefficients one can compute the calibration coefficients $c(v, s)$ using the update rule in Equation 12, which described the *mass measurement error* (*MME*) between the PL v and the starting PL s . The calibration is not accomplished until the whole tree is built. We then added PL v to the tree T and have iterated the procedure until all PLs were appended to the tree¹ (Figure 1 D).

In the MST algorithm, the vertices are joined by edges of smallest dissimilarity. Consequently, the MST algorithm connects all PLs in the data set in the way that the length of the path from the PL of origin (root of the tree: PL 3 in Figure 1 D) to any PL in the data set is minimal. The algorithm for computing the agglomerative clustering using the single linkage method [44, 45] works similarly like the MST algorithm and therefore the dendrogram (Figure 1 C) provides (as read from bottom to top) the order, by which the PL pairs were chosen (Figure 1 C). The horizontal lines joining two dendrogram tree branches were drawn at the height of the value of the minimal dissimilarity of two PLs in either branch.

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

Figure 2 A demonstrates how the samples on the target are connected by the edges. Green dots (brighter) represent leaves², while blue dots (darker) denote *interior* vertices. The PL of origin s is marked with a red cross-hairs (sample position D15). Note, that long PLs (brighter squares) are *interior* vertices of the MST. The strip-charts of mass ranges including peaks of the trypsin autolysis products 842.508, 2211.100 are given in Figure 2 C_1 and C_2 . One can observe that the MST-method works robustly on raw data with a mass measurement error of up to $\pm 0.7Da$ (black crosses), even if the search for matching peaks when computing the similarities and calibration coefficients was performed within a much smaller window of $\pm 0.45Da$. Notably, if the maximal error among two PLs is much larger than the search window, it is still possible to find a path, thus allowing alignment of two extreme PLs.

Due to the fact that all PLs were aligned to the PL 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 search (not shown).

Determining the calibration model of the sample support using Thin-Plate Spline interpolation (TPS)

Because a large part of the MME 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. [30]

¹ *e.g.*, by adding PL 4 then 5 and finally 1 to T (Figure 1)

² A leaf of a tree is a vertex with degree at most one. A vertex that is not a leaf is called an *interior* vertex.

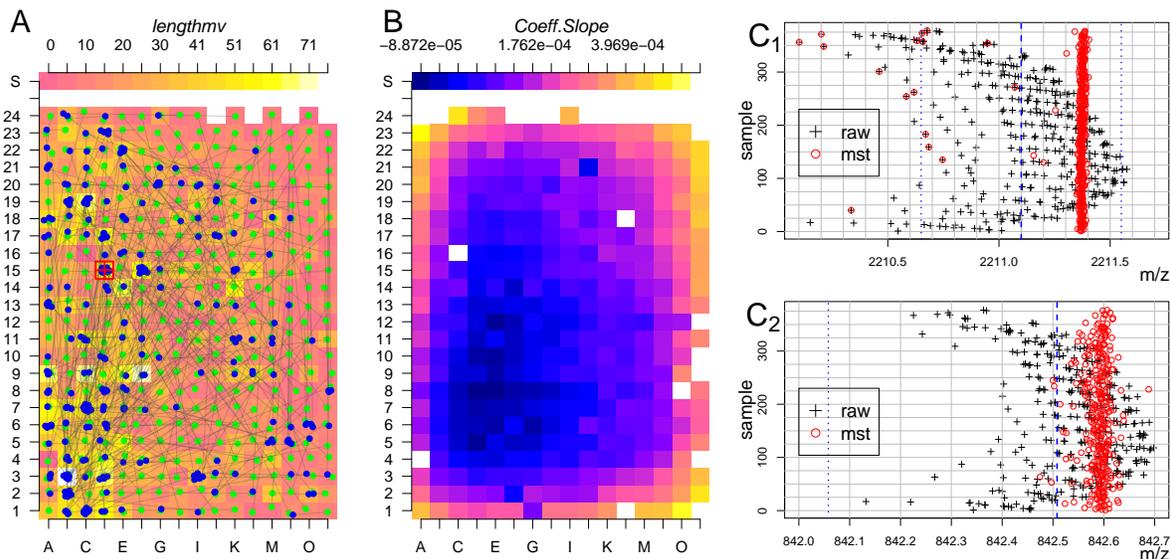


Figure 2: Figure **A**: BPY-colour scheme coded PL lengths in dependence of the sample support position. Blue dots – *interior* vertex, Green dots – end vertex, white arrows – connecting edges of the *MST*. The red hair-cross indicates the PL of origin s . Figure **B**: BPY-colour scheme coded slope coefficient of the mass-dependent calibration function in relation to sample support position. Figure C_1 , C_2 : Strip chart of the data set for a mass range of 2210 – 2212 Da (top) and 842 – 843 Da (bottom), including the tryptic autolysis peaks 842.508 and 2211.100 Da . 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 2211.100 Da .

and Moskovets et al. [31]. 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 afterwards measured on the same plate.

We introduced here a method that derives the calibration model from calibration coefficients obtained from MS analytes, which do not necessarily contain internal standards. Instead of refining *e.g.* the *MST* calibration model, we chose the peptide mass rule based approach, namely Linear Regression on Peptide Rule (LR/PR) (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 LR/PR method calibrates the PLs into the theoretical co-ordinate system and increases the mass accuracy to approximately 0.1 Da , but fails if the PL is too short, which indeed could be observed for several samples (Figure 3 A and C). Figure 3 A provides the colour scheme coded slope coefficient c_1 as determined by the LR/PR-method in dependence of the target location. One can observe that some erroneous predictions occur (Figure 3 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 2 B,

demonstrates that the slope coefficient of the sample calibration-model obtained by the MST calibration methods increases for samples close to the support border. This change is due to alterations in the electric field E (Equation 1) influencing the flight velocity given by

$$u_D = \sqrt{2Es_a \frac{z}{m}}, \quad (1)$$

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*(TPS) interpolation method [35,46]. At first, we computed the TPS with a degree of smoothing $\lambda = 5 \cdot 10^{-2}$ (see Equation 17). 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 estimated by the TPS, were discarded. Using the remaining calibration models, the TPS was recomputed with smaller degree of smoothing $\lambda = 1 \cdot 10^{-3}$. Figure 3 B, demonstrates the BPY-colour scheme coded slope coefficient c_1 , as estimated by the refined TPS. This model resembles the one generated by the MST method (Figure 2 B). We corrected the PLs masses (black cross hairs, Figure 3 C), using the TPS 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 TPS method reduced the MME of a PL compared to any other PL in the data set (vertical red, dashed line in Figure 3 C) down to $0.3Da$, as compared to $1.5Da$ for raw data. This is an ≈ 5 - fold increase of a mass measurement accuracy. This decrease of the MME enabled to utilise the MST-algorithm with an accuracy of $\pm 0.15Da$, reducing further the probability of false assignments of calibration masses. In addition, the histogram of dissimilarities computed for all PL pairs (Figure 4, A) shows for TPS 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 $0.15Da$ and the second with a search window of $0.45Da$. A subsequent calibration using the MST method further decreased the MME (Figure 4 B).

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 MME , 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 1 row 1 and 2). After removal of this error using calibration methods, *e.g.* TPS calibration (Table 1 row 3,4) or TPS with subsequent MST calibration

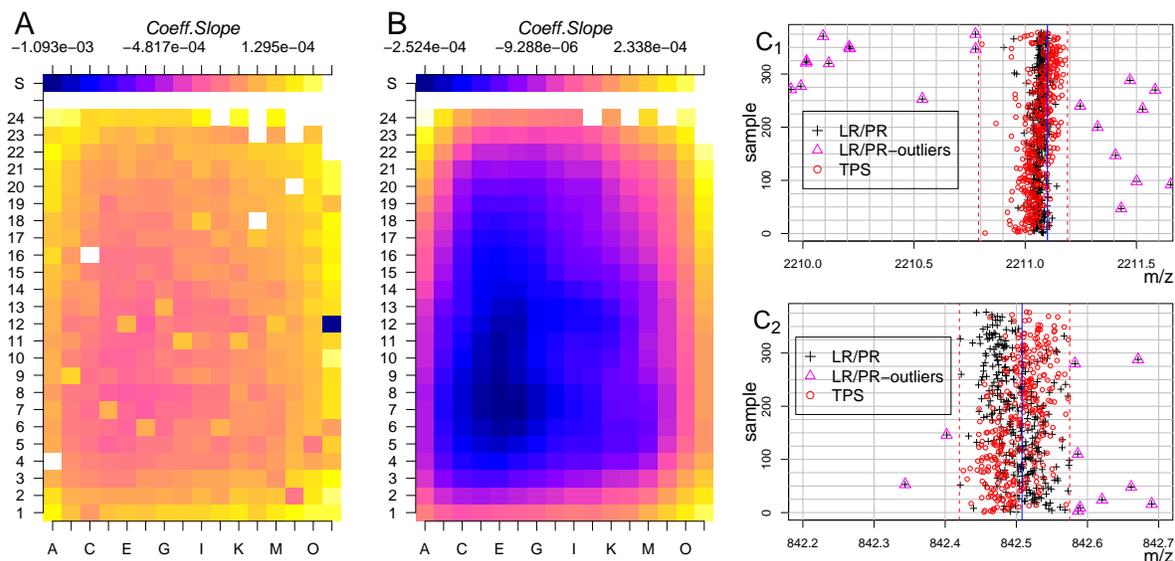


Figure 3: Figure **A** : BPY - colour scheme coded slope coefficients c_1 of the MME determined by LR/PR -method. Figure **B**: The slope coefficient as predicted from the refined samples determined by TPS with $\lambda = 0.001$. Figure **C**: Strip chart of the data set for a mass range of $2210 - 2212Da$ (C_1) and $842 - 843Da$ (C_2), including the tryptic autolysis peaks 842.508 and $2211.100Da$. Black crosses – masses of peaks predicted by LR/PR -calibration method, red circles – masses predicted by the TPS calibration method. Vertical blue line – exact position of trypsin autolysis masses 842.508 and $2211.100Da$. Dashed red vertical line – mass of the extreme peptide masses after TPS calibration.

(Table 1 row 5,6), the main contribution to the MME was due to peak detection performance³. The change of peak-detection quality was negligible in the range of $500 - 4000Da$. Figure 5, as well as Table 1 illustrates that after calibration the absolute MME 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 .

The optimal size of the search window

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 [48] search algorithm. The search window of $0.2Da$ generated the highest identification rate. Figure 6 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,

³We were aware however, of systematic changes of the MME , which can be described using higher order polynomials [47]. We have removed higher order terms of the MME , by applying external calibration prior to other calibration procedures (see Methods : Data Set).

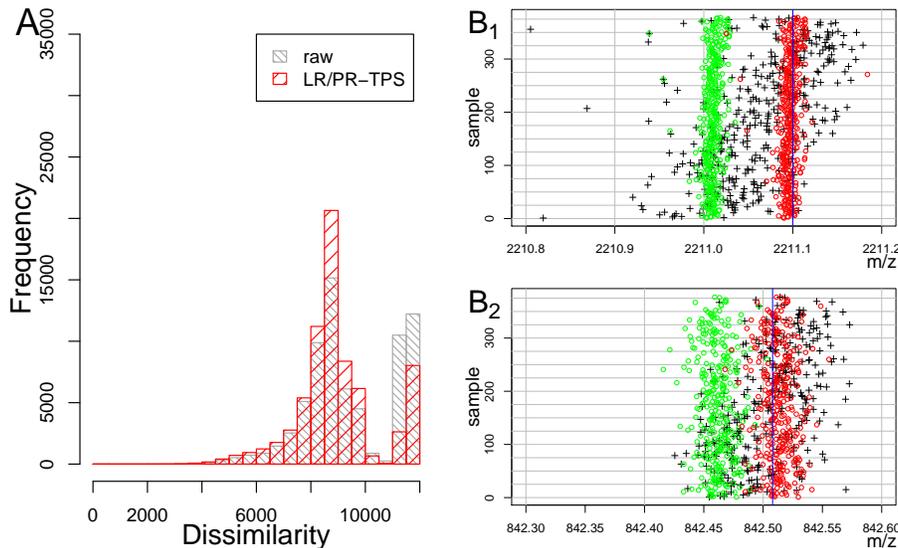


Figure 4: Figure **A**: Histogram of pairwise PL’s similarities. In gray – raw data and similarities computed with an accuracy of $\pm 0.4Da$. In red – similarities computed with accuracy of $\pm 0.15Da$ using LR/PR-TPS calibrated data. Figure **B**: Strip chart of PLs. Grey triangles – masses after TPS-calibration, green circles – data after *TPS-MST*-calibration, red circles – data calibrated into the theoretical co-ordinate system, defined by theoretical tryptic autolysis masses (blue vertical lines.)

while a smaller window *e.g.* $\pm 0.05Da$ decreases it by rejecting true matches [48]. Because the identification rate for a search window of $0.1Da$ is only slightly worse than one of $0.2Da$, and since it minimises the risk of false positive matches, we further compared the practical performance of the calibration methods with a search window of $0.1Da$.

Prior to the database searches we removed all masses that occur in more than 8% of spectra, as it significantly increased the identification rate [32,33] (cf Methods – Filtering of ubiquitous masses prior to database search). The sequence data base search was performed using the Mascot [48] search software version 1.8.1. We interfaced the search server from within R using the in-house developed R package *msmascot* [49].

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 [37] instrument. We applied the calibration methods introduced (MST and TPS 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

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 1: Mass Measurement Error. Standard deviation observed for the tryptic autolysis peaks 842.508 and 2211.1. Raw data; TPS - Thin-Plate Spline (TPS) calibrated data; TPS-MST - The data, which undergone Thin-Plate Spline (TPS)(pre-processing), followed by Maximum Spanning Tree (MST) calibration

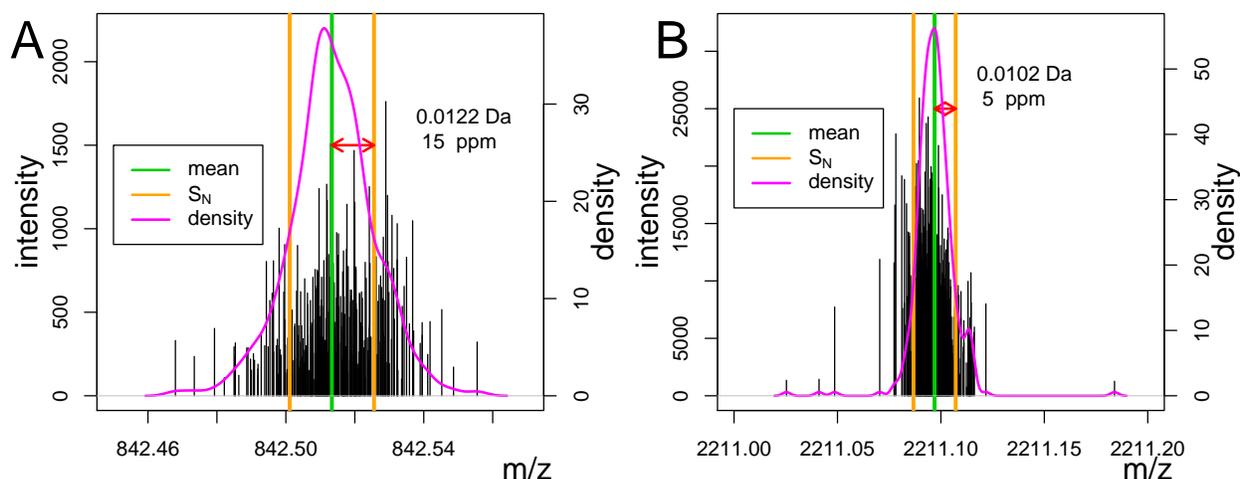


Figure 5: Stick Spectrum of the merged data set of 380 PLs. The black vertical lines represent peaks calibrated using the TPS and MST method. Their height equals their intensity. Green line – average mass of all peaks in the region 842 – 843Da (Figure A) and 2210.5 – 2211.6Da (Figure B). The orange vertical lines represent the average mass \pm , the standard deviation of the peak masses in each region. Magenta line – density of peak-masses.

optimised for the respective machines.

We combined the different pre-calibration and calibration methods resulting in six different calibration sequences, summarised in Table 2. We compared the performance of the MST and TPS 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 TPS based method could be improved further by subsequent internal (P-IC) or MST calibration (P-MST). The R [50] scripts implementing each sequence can be found in the samples directory of the `mscalib` BioConductor [51] package.

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

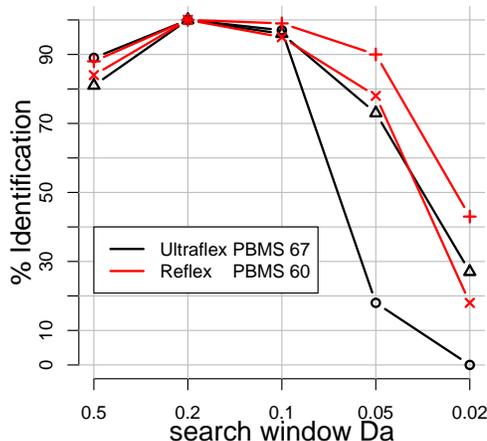


Figure 6: The optimal search window. Comparison of the relative identification rates of internally calibrated data (Y-axis) given a search window size of $0.5Da$, $0.2Da$, $0.1Da$, $0.05Da$ and $0.02Da$, respectively (X-axis). Red – Two Reflex (Pirellula) dataset, Black – Two Ultraflex (*Mus Musculus*) datasets

	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	P-IC	TPS calibration and subsequent internal calibration.
6	P-MST	TPS calibration and subsequent MST calibration.

Table 2: Calibration sequences. LR/PR – linear regression on peptide role, IC – Internal calibration with two iterations. (Bruker Reflex – mass measurement error (MME) window of 450 and 250ppm, Bruker Ultraflex – 250 and 125ppm); MST – MST calibration method computed with an search window of $\pm 0.4Da$; P-IC - Pre-processing (TPS calibration) and subsequent internal calibration with a MME window of 250ppm; P-MST - pre-processing and an MST with a search window of $\pm 0.25Da$;

trypsin autolysis peaks and a calibrated set of ubiquitous masses (cf Methods – Standard internal calibration). In case of the LR/PR calibration 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 PLs we performed the filtering.

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

$$\frac{\#\{CS_1 \cup CS_2 \cup \dots \cup CS_6\}}{\text{number of samples submitted for search}}, \tag{2}$$

where CS_i indicates the set of identified samples by one of the calibration sequences (Table 2), and $\#\{A\}$

denotes the number of elements in a set A, were in case of a) the Pirellula (Reflex) data set 74%, 87%, 79%, 85%, with an overall identification rate of 82% whereas b) in case of the *Mus musculus* (Ultraflex) data set: 51%, 72%, 35%, 51%, 27%, with an overall identification rate of 58%, respectively. 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):

$$\frac{\#\{CS_i\}}{\#\{CS_1 \cup CS_2 \cup \dots \cup CS_6\}} \quad (3)$$

The relative identification rate is indicated by the dots, joined by continuous lines for readability purposes only, in Figure 7. The dashed lines denote the average of the sequence coverage of all identified samples⁴. Figure 7A presents the results for the four Pirellula data sets, while Figure 7B shows the results of five *Mus musculus* data sets.

Only in one case of one data set was a single calibration sequence P-MST (see Table 2) able to identify all PLs (100% identification rate) and therefore it completely dominated over the other methods (black line, Figure 7 A). In case of the Ultraflex data set (Figure 7 B) we observed that the P-MST method had the highest identification rate, while in Reflex data set (Figure 7 A) it achieved the highest performance for approximately half of the data sets.

Figure 7 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 [48] (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 TPS smoothing method resulted in an overall higher identification rate than the other methods tested on raw data (LR/PR, IC, MST), except for one case of the Ultraflex data set.

⁴We compute the average sequence coverage for all spectra that were identified by at least one of the methods.

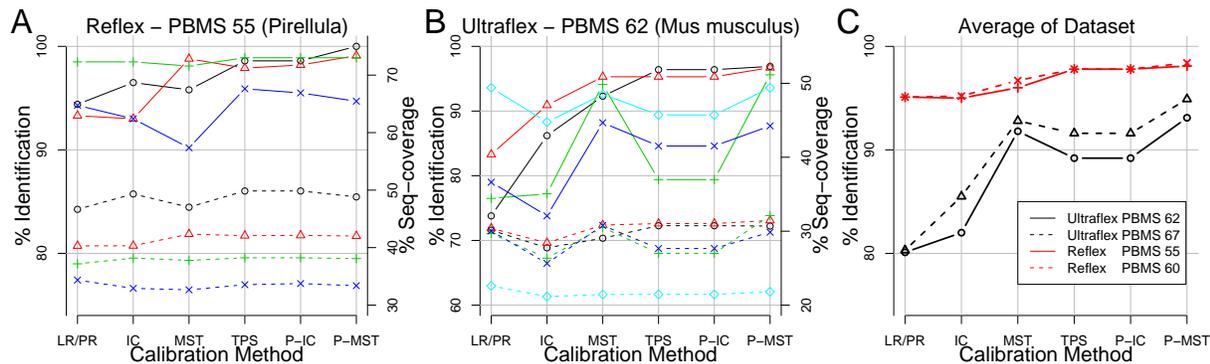


Figure 7: Relative identification rate in % (continuous line – left y-axis) and sequence coverage in % (dashed lines - right y-axis). LR/PR – linear regression on peptide rule, IC – two step internal calibration, MST – minimum spanning tree calibration, P – TPS calibration, P-IC – TPS calibration and subsequent internal calibration, P-MST - TPS calibration and subsequent MST calibration.

Furthermore, a combination of the internal calibration with TPS calibration (P-IC) did not increase either the sequence coverage (dashed lines) or the identification rate of the TPS method applied alone.

In two out of the four Reflex data sets the MST method applied on TPS-processed data (P-TPS Figure 7 A, dashed lines) slightly decreased the sequence coverage indicating a reduction of calibration accuracy. In case of the Ultraflex data sets, the sequence coverage correlated well with the identification rate and the P-MST-method accomplished the highest performance.

Moreover, if similar identification rates of the LR/PR and the IC method were observed, the LR/PR method provided higher sequence coverage (Figure 7 B). This could be explained by the fact that the LR/PR method calibrated well the PLs possessing many peptide peaks. Such PLs potentially contain the higher sequence coverage.

The BioConductor package `mscalib`

All of the calibration methods are part of the `mscalib` programme, which is available as a *BioConductor* [52] package. The Bioconductor project is an initiative for the collaborative creation of extensible software for computational biology and bioinformatics [51]. 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 [53] with detailed descriptions of two selected calibration sequences.

Conclusions

Usage of TPS calibration results in up to 10% higher identification rates than the internal calibration. The TPS 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 the peptide mass rule. The MST method is able to increase the identification rates obtained by the TPS-method. 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 MST algorithm that builds a single tree for all PLs. This is adequate if the data are a set of PLs with smooth transitions in the similarity values. If this is not the case, it might be more appropriate to compute a forest of several MSTs. We have examined, however, only a single PL similarity measure Equation 11 for PLs calibration. It is possible that better similarity measures can still be generated and subsequently applied for PLs calibration.

Complete utilisation of microtitre plates and sample supports is not only rational with respect to increased accuracy of the TPS method, but also with respect to the idea of HT experiments – maximal utilisation of energy and resources. Dense excision of spots from 2D-gels not only increases the performance of the MST method, but also identifies novel proteins. Hence, the main contribution of this manuscript is to present two calibration methods, compatible with the principle of HT sample processing and aims to identify a maximum of the proteins resolved on 2D-gels.

However, there is no single best calibration method. Each of the methods utilises different properties of the PLs. Consequently, applying these methods in parallel and determining the total of identified samples provides the highest identification rate.

Methods

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 [37] MALDI-TOF instrument.
2. A mammalian proteome *Mus musculus* (1,882 spectra) measured on Ultraflex [37] MALDI-TOF instrument.
3. A plant proteome *Arabidopsis thaliana* [36] measured on an Autoflex [37] 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 [3] and visualised with MS compatible Coomassie brilliant blue G250 [36]. The MALDI-TOF MS analysis was performed using delayed ion extraction and by employing the MALDI AnchorChip TMtargets (Bruker Daltonics, Bremen, Germany). Positively charged ions in the m/z 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 (PL). Processed PLs were then used for the protein database searches with the Mascot search software (Version 1.8.1) [48], 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.

Describing the Mass Measurement Error (MME) and predicting the correct mass

A mass difference can be described either in absolute $\Delta_A = m_y - m_x[m/z]$ or in relative $\Delta_R = (m_y - m_x) \cdot 10^6/m_y[ppm]$ units. The masses in two PL's X, Y were compared to each other and we considered two peaks to *match*, in case of the absolute error if $\Delta_A < a[m/z]$ and in 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 $\epsilon \in N(0, \sigma)$, a systematic dependence. This dependence we modelled using a function $\hat{f}(m)$. Given $\hat{f}(m)$ we corrected the experimental masses using equations:

$$m_{corr} = \frac{m_{exp}}{1 - \hat{f}_R(m_{exp}) \cdot 1/10^6}, \text{ or} \quad (4)$$

$$m_{corr} = m_{exp} + \hat{f}_A(m_{exp}), \quad (5)$$

in case of the relative or absolute error respectively, to attain corrected masses m_{corr} .

Affine MME model

In the first approximation, the *MME* 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 peaks⁵. The intercept and slope coefficients of this function can be determined using linear regression.

⁵Because the mass m is much larger than Δ , in practice it does not matter whether we choose the average mass of the matching peaks, the theoretical mass m_{theo} , or the experimental mass m_{exp}

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, in case of absolute $\Delta_A[Da]$
- the slope coefficient at 0, in case of relative $\Delta_R[ppm]$

and determine the slope or intercept respectively from the data.

To correct the experimental masses m_{exp} we used Equation 5 for the absolute differences Δ_A of matching peaks and Equation 4 in case of relative differences Δ_R .

The difference between theoretical and measured masses is called mass measurement error *MME*, while the alignment of m_{exp} on m_{theo} an *internal calibration* [17, 47, 54].

Determining ubiquitous masses and their filtering

To determine the abundant masses we computed two histograms for each data set. The origin of the first histogram \hat{f}_h^1 is $x_0 = \min(M) - h$ and of the second \hat{f}_h^2 $x_0 = \min(M) - h/2$, 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

$$B_j = [x_0 + (j - 1)h, x_0 + jh], \quad \text{with } j \in 1, \dots, l, \quad (6)$$

where $l = (\max(M) - x_0) \text{ mod } h$. Formally the histogram of counts f is given by [55]

$$\hat{f}_h(x) = \sum_{i=1}^n \sum_j^l I(X_i \in B_j)I(x \in B_j), \quad (7)$$

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

$$m = \frac{m_1 \cdot c_1 + m_2 \cdot c_2}{c_1 + c_2}, \quad (8)$$

where m_1 and m_2 are the bin mid. 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. Usage of 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. [33] and Kreitler [56].

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 PLs in the data set were then aligned to the calibrated list of ubiquitous masses.

Filtering of ubiquitous masses prior to database search.

We removed ubiquitous masses that occurred in more than 7.7% of PLs [32,33]⁶. Filtering of ubiquitous masses was performed on a calibrated set of PLs. As a result, we could use a small bandwidth of $h = 0.2Da$ (Equation 6) 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 ⁷. Finally, we removed all peaks within the range $\pm 0.1Da$ around the ubiquitous masses.

Linear Regression and Peptide mass Rule - LR/PR algorithm.

Wolski et al. (publication in preparation) defined the distance measure

$$d_{\lambda}(m_i, m_j) = \begin{cases} |m_i - m_j| \bmod \lambda_{DB} & \text{if } |m_i - m_j| \bmod \lambda_{DB} < 0.5 \\ -(1 - |m_i - m_j| \bmod \lambda_{DB}) & \text{if } |m_i - m_j| \bmod \lambda_{DB} \geq 0.5 \end{cases}, \quad (9)$$

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 [41]. 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 *MME*. 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 [57]. In order to make the prediction robust against *e.g.* non-peptide peaks, we used a robust linear regression [58]. We removed the

⁶Levander et al. suggested to use as a threshold the function considering the distribution of peptides in the protein database normalised by the mass range and the number of peptides in the data set. The package `mscalib` implements this function. We did not use it because in our experience the frequency distribution of peptides in the data set depends much more on the mass-spectrometric matrix used than it is influenced by the theoretical peptide distribution determined for a database.

⁷Abundant 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

slope by multiplying each mass m_i in the PL 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.

Higher order Calibration Function

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 [59]. To model the dependence $\Delta \propto m$ we used a cubic smoothing spline function [60, 61], 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 [62] 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 [30], which yield equidistant peaks were used to precisely estimate the mass-dependent difference function.

Similarity/Quality measures for internal calibration

PLs 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 PL pair (X, Y) fails if no matching peaks are found. We described these properties mathematically by the following similarity measures:

$$S_{X,Y} = \left(\sum_{i=1}^n \sum_{j=i+1}^n (|m_j - m_i|)^p \right)^{1/p}, \text{ and} \quad (10)$$

$$F_{X,Y} = \frac{S_{A,B}}{\frac{\min(|X|, |Y|) \cdot (\min(|X|, |Y|) - 1)}{2} \cdot (\max(X \cup Y) - \min(X \cup Y))}, \quad (11)$$

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. To normalise $S_{X,Y}$, we divided the similarity by the product of the maximal possible number of matching peaks ($\min(|X|, |Y|)$) times the mass range bracketed by the peaks of both PLs. Consequently, $F_{X,Y} \in [0, 1]$ since the denominator is always larger or equal to $S_{X,Y}$.

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 $v \in V \setminus U$ 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 PL. 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 PLs (vertices) to the starting PL (vertex) s .
- 12 W – which are the weights of the path from $s \rightarrow v \in F$.
- 13 S – modified similarity matrix.

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

Alignment of a set of PL using a Minimum Spanning Tree

To align a whole data-set to a single PL and to align the PLs with the highest similarity given by Equation 10, first we computed for all PLs 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 PLs 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(MST) [43] was implemented. We present here a modified version of this algorithm (see Figure 8). 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 MST tree T , the algorithm returns also a list of

calibration coefficients C , which align all PLs V in the data set to the starting vertex (PL) 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 8, line 5).

To align PL v to the starting PL s we needed to determine the coefficients $C(v, s)$ of the difference function $\hat{f}(v, s)$ (Equation 5). 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:

$$C(v, s) = C(v, u) \circ C(u, s) = \begin{cases} c_1^{vs} & = c_1^{vu} + c_1^{us} + c_1^{vu}c_1^{us} \\ c_0^{vs} & = c_0^{us} + c_0^{vu} + c_1^{vu}c_0^{us} \end{cases}, \quad (12)$$

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

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

$$\begin{aligned} m_u &= m_v + \hat{f}_A(v, u) = m_v + c_1^{vu} \cdot m_v + c_0^{vu}, \text{ and} \\ m_s &= m_u + \hat{f}_A(u, s) = m_u + c_1^{us} \cdot m_u + c_0^{us} \text{ (see Equation 5)}. \end{aligned}$$

Hence,

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

$C(v, s)$ was computed online using Equation 12, while growing the tree (Figure 8, line 8). Subsequently, the algorithm returned a list C of calibration constants, where $C(v, s)$ described the calibration coefficients allowing to transform PL v into the co-ordinate system of the PL of origin s .

In order to gain more confidence in the calibration constants in C , the MST algorithm was iterated n times. For computing the consecutive T_i, C_i, W_i, D_i with $i = 2, \dots, n$ 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 8, line 6).

The calibration constants $C_i(v, s)$ with $i = 1, \dots, n$ 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 8, line 9), on the path from s to v .

$$C_w(v, s) = \sum_{i=1}^n \frac{W_i(v) \cdot C_i(v, s)}{W_i(v)}. \quad (13)$$

We applied the calibration constants in C_w to align all PLs to the PLs s_1 .

Appendix

Smoothing spline function

To model the dependence $y_i \sim x_i$ we used a cubic smoothing spline function [55, 60, 61]. Specifically, we let $y_i = f(x_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ and f was a smooth function estimated by

$$\hat{f}(x) = \operatorname{argmin}_{f \in C^2[x_1, x_n]} \left(\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_{x_1}^{x_n} (f''(x))^2 dx \right) \quad (14)$$

The first term measures the closeness to the data, while the second penalises curvatures in the function and λ established a trade off between the two. If $\lambda \rightarrow \infty$, then $\hat{f}(x)$ becomes a linear function. As $\lambda \rightarrow 0$, $\hat{f}(x)$ is merely an cubic interpolating spline of the observations of Y . The unique solution of Equation 14 is a natural cubic spline (NCS).

A different way to specify the degree of smoothing is by fixing the *effective degrees of freedom*, which equal the trace of the smoother matrix $\operatorname{trace}(S_\lambda) = df_\lambda$, where S_λ is $\hat{f}(x_i) = S_\lambda y_i$. Since $df_\lambda = \operatorname{trace}(S_\lambda)$ is monotone in λ for smoothing splines. Thus, we could specify λ by fixing the degrees of freedom [63].

Thin-plate spline

The thin-plate spline is the two-dimensional analogue to the cubic spline in one dimension [35, 64]. 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

$$f(x, y) = a_1 + a_x x + a_y y + \sum_{i=0}^p w_i U(\|(x_i, y_i) - (x, y)\|), \quad (15)$$

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 [35, 64] which minimises the equation:

$$I[f(x, y)] = \int \int_{R^2} \left(\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right) dx dy \quad (16)$$

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 16). This can be accomplished by multiplying equation 16 with a *regularization* parameter λ , a positive scalar, and by adding the residual sum of squares (*RSS*), which gives:

$$H[f(x, y)] = \sum_{i=1}^n (v_i - f(x_i - y_i))^2 + \lambda \cdot I[f(x, y)] \quad (17)$$

Again, as in case of the cubic smoothing spline by the parameter λ , the degree of smoothing can be determined. In our study, we utilised an implementation of the TPS [65], according to Doug Nychka [46].

Agglomerative Clustering

At the beginning of the clustering algorithm we had a $P^n = v_1, \dots, v_n$ clustering solution, where each PL v_i in the data set belonged to a separate cluster C_i . We iteratively constructed a solution P^k from P^{k+1} , by choosing two clusters C_h and C_l from P^{k+1} . We removed C_h and C_l and added the merged cluster $C = C_h \cup C_l$. The clusters C_h and C_l were selected according to the chosen linkage method. Single linkage defines the distance between any two clusters as the minimum distance between them (Equation 18), *i.e.* the distance between the two closest points (entities) C_l and C_h , given by

$$d_{\text{single}}(C_l, C_h) = \min_{x \in C_h, y \in C_l} d(x, y) \quad (18)$$

$$d_{\text{complete}}(C_l, C_h) = \max_{x \in C_h, y \in C_l} d(x, y) . \quad (19)$$

Usage of a single linkage method often causes the chaining phenomenon: forcing clusters together due to single entities being close to each other, regardless of the positions of other entities in the cluster. Equation 19 defines the distance for the complete linkage method. Consequently, the complete linkage minimised the diameter of the new cluster $C = C_h \cup C_l$.

Abbreviation

- MME - mass measurement error
- HT - high throughput.
- BPY - blue-pink-yellow colour scheme that prints well as grey tone.

- *MST* - minimum spanning tree.
- MS - Mass Spectrometry.
- TOF - Time of Flight.
- MALDI - Matrix Assisted Laser Desorption Ionization.
- mod - modulo operator.
- RSS - residual sum of squares.
- PL - peak-list

Authors contributions

ML and PJ gave initial input to the research.

WEW implemented the *BioConductor* package *mscalib,msmascot*, carried out the analysis, visualised the results and wrote the manuscript.

ML wrote essential parts of the manuscript

All authors contributed to the final version of the manuscript and approved it.

Acknowledgements

We would like to thank the members of Algorithmic Bioinformatics group at FU-Berlin for valuable discussion, especially Dr. Clemens Gröpl. We would like to thank Dr. Johan Gobom, Dr. Patrick Giavalisco and Thomas Kreitler for providing the PMF-MS data and for valuable discussion. We thank Carole Procter, Stale Nygard, Dan Swan and Daniel Henderson for proofreading the manuscript. We thank Prof. Dr. Hans Lehrach at which department part of the work was performed. This project was funded by the National Genome Research Network (NGFN) of the German Ministry for Education and Research (BMBF), and the Max Planck Society.

References

1. Gevaert K, Vandekerckhove J: **Protein identification methods in proteomics.** *Electrophoresis* 2000, **21**(6):1145–54.
2. Blackstock W, Weir M: **Proteomics: quantitative and physical mapping of cellular proteins.** *Trends Biotech* 1999, **17**:121–127.

3. Klose J, Kobalz U: **Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome.** *Electrophoresis* 1995, **16**(6):1034–59.
4. Quadroni M, James P: **Proteomics and automation.** *Electrophoresis* 1999, **20**:664–677.
5. Nordhoff E, Egelhofer V, Giavalisco P, Eickhoff H, Horn M, Przewieslik T, Theiss D, Schneider U, Lehrach H, Gobom J: **Large-gel two-dimensional electrophoresis-matrix assisted laser desorption/ionization-time of flight-mass spectrometry: an analytical challenge for studying complex protein mixtures.** *Electrophoresis* 2001, **22**(14):2844–2855. [(eng)].
6. Fenyo D: **Identifying the proteome: software tools.** *Current Opinion in Biotechnology* 2000, **11**:391–395.
7. Griffin TJ, Aebersold R: **Advances in proteome analysis by mass spectrometry.** *J Biol Chem* 2001, **276**:45497–500.
8. Patterson SD: **Data analysis—the Achilles heel of proteomics.** *Nat Biotechnol* 2003, **21**(3):221–2.
9. Aebersold R, Mann M: **Mass spectrometry-based proteomics.** *Nature* 2003, **422**(6928):198–207.
10. Mann M, Hojrup P, Roepstorff P: **Use of mass spectrometric molecular weight information to identify proteins in sequence databases.** *Biol Mass Spectrom* 1993, **22**(6):338–345.
11. Pappin DJC, Hojrup P, Bleasby AJ: **Rapid identification of proteins by peptide-mass fingerprinting.** *Curr. Biol.* 1993, **3**:327–332.
12. Colby SM, King TB, Reilly JP: **Improving the Resolution of MALDI TOF Mass Spectrometry by Exploiting the Correlation Between Ion Position and Velocity.** *Rapid Comm. Mass Spectrom.* 1994, **8**:865–868.
13. Whittall RM, Li L: **High-resolution matrix-assisted laser desorption/ionization in a linear time-of-flight mass spectrometer.** *Anal Chem* 1995, **67**(13):1950–4.
14. Brown RS, Lennon JJ: **Mass resolution improvement by incorporation of pulsed ion extraction in a matrix-assisted laser desorption/ionization linear time-of-flight mass spectrometer.** *Anal Chem* 1995, **67**(13):1998–2003.

15. Takach EJ, Hines WM, Patterson DH, Juhasz P, Falick AM, Vestal ML, Martin SA: **Accurate mass measurements using MALDI-TOF with delayed extraction.** *J Protein Chem* 1997, **16**(5):363–9.
16. Guilhaus M: **Principles and Instrumentation in Time-of-Flight Mass Spectrometry.** *JOURNAL OF MASS SPECTROMETRY* 1995, **30**:1519–1532.
17. Gras R, Muller M, Gasteiger E, Gay S, Binz PA, Bienvenut W, Hoogland C, Sanchez JC, Bairoch A, Hochstrasser DF, Appel RD: **Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection.** *Electrophoresis* 1999, **20**(18):3535–3550. [(eng)].
18. Wool A, Smilansky Z: **Precalibration of matrix-assisted laser desorption/ionization-time of flight spectra for peptide mass fingerprinting.** *Proteomics* 2002, **2**(10):1365–1373.
19. Strittmatter EF, Rodriguez N, Smith RD: **High mass measurement accuracy determination for proteomics using multivariate regression fitting: application to electrospray ionization time-of-flight mass spectrometry.** *Anal Chem* 2003, **75**(3):460–8.
20. Samuelsson J, Dalevi D, Levander F, Rognvaldsson T: **Modular, Scriptable, and Automated Analysis Tools for High-Throughput Peptide Mass Fingerprinting.** *Bioinformatics* 2004.
21. Apweiler R, Bairoch A, Wu CH: **Protein sequence databases.** *Curr Opin Chem Biol* 2004, **8**:76–80.
22. Pappin J, Hojrup P, Bleasby A: **Rapid Identification of Proteins by Peptide-Mass Fingerprinting.** *Current Biology* 1993, **3**:327–332.
23. Zhang W, Chait BT: **ProFound: an expert system for protein identification using mass spectrometric peptide mapping information.** *Anal Chem* 2000, **72**(11):2482–2489.
24. Eriksson J, Fenyo D: **A Model of random mass-matching and its use for automated significance testing in mass spectrometric proteome analysis.** *Proteomics* 2002, **2**(3):262–70.
25. Parker KC: **Scoring methods in MALDI peptide mass fingerprinting: ChemScore, and the ChemApplex program.** *J Am Soc Mass Spectrom* 2002, **13**:22–39.

26. Tabb DL, Huang Y, Wysocki VH, Yates JRr: **Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides.** *Anal Chem* 2004, **76**(5):1243–8.
27. Pevzner PA, Dancik V, Tang CL: **Mutation-Tolerant Protein Identification by Mass Spectrometry.** *Journal of Computational Biology* 2000, **7**(6):777–787.
28. Egelhofer V, Gobom J, Seitz H, Giavalisco P, Lehrach H, Nordhoff E: **Protein identification by MALDI-TOF-MS peptide mapping: A new strategy.** *Analytical Chemistry* 2002, **74**(8):1760–1771.
29. Schuerenberg M, Luebbert C, Eickhoff H, Kalkum M, Lehrach H, Nordhoff E: **Prestructured MALDI-MS sample supports.** *Anal Chem* 2000, **72**(15):3436–42.
30. Gobom J, Mueller M, Egelhofer V, Theiss D, Lehrach H, Nordhoff E: **A calibration method that simplifies and improves accurate determination of peptide molecular masses by MALDI-TOF MS.** *Anal Chem* 2002, **74**(15):3915–3923. [(eng)].
31. Moskovets E, Chen HS, Pashkova A, Rejtar T, Andreev V, Karger BL: **Closely spaced external standard: a universal method of achieving 5 ppm mass accuracy over the entire MALDI plate in axial matrix-assisted laser desorption/ionization time-of-flight mass spectrometry.** *Rapid Commun Mass Spectrom* 2003, **17**(19):2177–87.
32. Chamrad DC, Koerting G, Gobom J, Thiele H, Klose J, Meyer HE, Blueggel M: **Interpretation of mass spectrometry data for high-throughput proteomics.** *Anal Bioanal Chem* 2003, **376**(7):1014–22.
33. Levander F, Rognvaldsson T, Samuelsson J, James P: **Automated methods for improved protein identification by peptide mass fingerprinting.** *Proteomics* 2004, **4**(9):2594–601.
34. Wolski WE, Lalowski M, Martus P, Herwig R, Giavalisco P, Sickmann A, Lehrach H, Gobom J, Reinert K: **Transformation and other factors of the biological Mass Spectrometry pairwise peak-list Comparison Process.** *BMC Bioinformatics* 2005.
35. Bookstein F: **Principal Warps: Thin-Plate Splines and the Decomposition of Deformations.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1989, **11**(6):567–585.

36. Giavalisco P, Nordhoff E, Kreitler T, Kloeppe KD, Lehrach H, Klose J, Gobom J: **Proteome Analysis of Arabidopsis Thaliana by 2-D Electrophoresis and Matrix Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometry**. [To appear in Proteomics].
37. **Bruker Daltonics - enabling life science tools based on mass spectrometry**. 2004, [<http://www.bdal.com>].
38. Glockner FO, Kube M, Bauer M, Teeling H, Lombardot T, Ludwig W, Gade D, Beck A, Borzym K, Heitmann K, Rabus R, Schlesner H, Amann R, Reinhardt R: **Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1**. *Proc Natl Acad Sci U S A* 2003, **100**(14):8298–303.
39. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence project: update and current status**. *Nucleic Acids Res* 2003, **31**:34–7.
40. Thiede B, Lamer S, Mattow J, Siejak F, Dimmler C, Rudel T, Jungblut PR: **Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion**. *Rapid Commun Mass Spectrom* 2000, **14**(6):496–502.
41. Gay S, Binz PA, Hochstrasser DF, Appel RD: **Modeling peptide mass fingerprinting data using the atomic composition of peptides**. *Electrophoresis* 1999, **20**(18):3527–3534. [(eng)].
42. Schmidt F, Schmid M, Jungblut PR, Mattow J, Facius A, Pleissner KP: **Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis**. *J Am Soc Mass Spectrom* 2003, **14**(9):943–56.
43. Schrijver A: *Combinatorial Optimization - Polyhedra and Efficiency*. Berlin: Springer-Verlag 2003.
44. Härdle W, Simar L: *Applied Multivariate Statistical Analysis*. Springer, Heidelberg 2003, [<http://www.quantlet.com/mdstat/scripts/mva/htmlbook/mvahtml.%html>].
45. Handl A: *Multivariate Analysemethoden - Theorie und Praxis multivariater Verfahren unter besonderer Berücksichtigung von S-PLUS*. Springer, Heidelberg 2003, [<http://www.quantlet.com/mdstat/scripts/mst/html>].
46. Nychka D: **fields - A collection of programs based in [R,S] for curve and function fitting with an emphasis on spatial data**. 2004, [<http://www.cgd.ucar.edu/stats/Software/Fields/>].

47. Gobom J, Mueller M, Egelhofer V, Theiss D, Lehrach H, Nordhoff E: **A Calibration Method that Simplifies and Improves Accurate Determination of Peptide Molecular Masses by MALDI-TOF-MS.** *Analytical Chemistry* 2002, **74**(8):3915–3923.
48. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20**(18):3551–3567.
49. **R for Proteomics**[<http://r4proteomics.sourceforge.net>].
50. R Development Core Team: **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria 2004, [<http://www.R-project.org>]. [ISBN 3-900051-00-3].
51. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Li FLC, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: Open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **5**:R80, [<http://genomebiology.com/2004/5/10/R80>].
52. **Bioconductor - open source software for bioinformatics** 2004, [<http://www.bioconductor.org>].
53. Leisch F: **Sweave and Beyond: Computations on Text Documents.** In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Edited by Hornik K, Leisch F, Zeileis A, Technische Universität Wien, Vienna, Austria 2003.
54. Lee K, Bae D, Lim D: **Evaluation of parameters in peptide mass fingerprinting for protein identification by MALDI-TOF mass spectrometry.** *Mol Cells* 2002, **13**(2):175–84.
55. Härdle W, Müller M, Sperlich S, Werwatz A: *Nonparametric and Semiparametric Models - An Introduction.* Springer, Heidelberg 2004, [<http://www.quantlet.com/mdstat/scripts/spm/html/spmhtml.html%>].
56. Kreitler T: **Oral Communication** 2003.
57. Chambers JM: **Linear models.** In *Statistical Models in S.* Edited by Chambers J, Hastie T, Wadsworth & Brooks/Cole 1992.

58. Venables WN, Ripley BD: *Modern Applied Statistics with S. Fourth Edition*. Springer 2002, [<http://www.stats.ox.ac.uk/pub/MASS4/>]. [ISBN 0-387-95457-0].
59. Gobom J, Schürenberg M, Mueller M, Theiss D, Lehrach H, Nordhoff E: **alpha-cyano-4-hydroxycinnamic acid affinity sample preparation. A protocol for MALDI-MS peptide analysis in proteomics**. *Analytical Chemistry* 2001, **73**(3):434–438.
60. Chambers JM, Hastie TJ: *Statistical Models in S*. London: Chapman & Hall 1992.
61. Hastie T, Tibshirani R: *Generalized Additive Models*. Chapman and Hall 1990.
62. Cleveland W, Grosse E, Shyu W: **Local Regression Models**. In *Statistical Models in S*. Edited by Chambers J, Hastie T, Wadsworth & Brooks/Cole 1992.
63. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning*. Springer 2001. [ISBN:0387952845].
64. Donato G, Belongie S: **Approximation Methods for Thin Plate Spline Mappings and Principal Warps**. In *Computer Vision - ECCV 2002: 7th European Conference on Computer Vision, Copenhagen, Denmark, May 28-31, 2002. Proceedings, Part III*, Lecture Notes in Computer Science. Edited by Heyden A, Sparr G, Nielsen M, Johansen P, Springer-Verlag Heidelberg 2002:21–31.
65. Green P, Silverman B: *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall 1994.