Chapter 3

A mathematical model of the peptide mass rule with applications

3.1 Introduction

The mass spectrometric technique is widely used to identify proteins in biological samples (6; 8; 9; 10). The proteins are cleaved into peptides by a residue specific protease, *e.g.* trypsin. The resulting cleavage products can then be analysed by peptide mass fingerprinting (18) or subjected to MS/MS fragment ion analysis (24; 25), which both rely on the comparison of peptide or peptide fragment ion spectra with spectra simulated from protein sequence databases (104).

The sensitivity and specificity of the peptide identification can be increased by various post-processing methods, for example calibration (30; 37; 40; 41) and identification of non-peptide peaks (30; 42; 43). The fact that peptide masses are not uniformly distributed across the mass range but form equidistantly spaced clusters (51) is employed by some of these methods. In dependence on the atomic composition of the peptide, the monoisotopic mass would emerge below (e.g. cystein rich peptides) or above (e.g. lysine rich peptides) the cluster centres. The deviation from the cluster centre is a result of the mass defect, which is the difference between the nominal mass and the monoisotopic mass (Table 3.1). The mass defect is a result of atom fusion (79; 105).

	Atom	monoisotopic	nominal	mass defect
1	H	1.00782	1	0.00782
2	C	12.00000	12	0.00000
3	N	14.003074	14	0.003074
4	O	15.99491	16	−0.00032
5	S	31.97207	32	−0.00087

Table 3.1: Masses of Atoms

Calibration Mass spectrometric peptide peak-lists of peptide mass finger print experiments (19) can be calibrated by comparing the location of measured peptide masses with the location of the peptide mass cluster centres. Gras et al. (35) suggested the use of maximum likelihood methods in order to determine the calibration coefficients a and b. They defined the likelihood function by:

$$\sum_i P(am_i + b, \Delta m) \ , \tag{3.1}$$

where m_i is the i-th mass in the peak-list, and Δm is a search window. $P(m, \Delta m)$ is the probability to find a mass in $[m, m + \Delta m]$ given the theoretical distribution of peptide masses. The parameters a, b for $\arg_{\max} \sum_i P(am_i+b, \Delta m)$ can then be used to calibrate the peak-lists. The authors, however, do not provide information on whether $P(m, \Delta m)$ was determined from the exact distribution of the peptide masses or if a model approximating the distribution was used. They also do not mention which algorithm was used to maximise the likelihood. They reported that a mass measurement accuracy of $0.2Da$ and better was obtained after calibration.

Wool and Smilansky (30) have used *Discrete Fourier Transformation* (DFT) to determine the frequency λ and phase φ of a peak-list or mass spectrum. By comparing the experimental λ and φ with the theoretical $\lambda = 1.000495$ and $\varphi = 0$, they determined the slope and intercept of the calibration function. The authors reported a $40-60\%$ reduction of the mass measurement error. Furthermore, they presented a scoring scheme for sequence database searches. This scoring scheme

approximates the probability $P(m, \Delta m)$ to observe a peptide peak of mass m with given measurement error Δm.

Matrix noise Filtration. The most widely used MALDI matrices for the analysis of peptides are 3,5-Dimethoxy-4-hydroxycinnamic acid (*synapic acid*), alpha-Cyano-4-hydroxycinnamic acid (*alpha cyano*) (70) and 2,5-dihydroxybenzoic acid (*DHB*) (52). Unfortunately, clusters of matrix molecules can be ionised and cause peaks in the same mass range where peptide peaks are measured. Matrix aggregate formation can be minimised but not eliminated by adding ammonium acetate (52).

Some of the database search scoring schemes incorporate the number of signals (peaks) not assigned to a protein when computing the identification scores (55). Therefore, the presence of matrix signals in MS spectra decreases the sensitivity of the MS spectra interpretation. Hence, the removal of peaks strongly deviating from the cluster centres is applied (52; 53). The measure of deviation from cluster centres introduced here provides a simple tool to filter non-peptide peaks.

Data Reduction A further application which employs the property of peptide mass clustering is the binning of the mass measurement range. By applying this technique the amount of data is reduced, thus increasing the speed with which the pairwise comparison of spectra can be made (5; 50).

All these applications require us to know the exact location of or the distance between the peptide mass cluster centres. The distance between the cluster centres, which we will henceforth call wavelength λ, is commonly computed by first generating an *in silico* digest of the database. Afterwards, the linear dependence between the decimal point and the integer part is determined by regression analysis, for a relatively small mass range of 500 to $1000Da$ (53). Various authors report different values of the distance between clusters: Wool and Smilansky reported 1.000495 (30), Gay et al. 1.000455 (51), while Tabb et al. used a wavelength of 1.00057 (5).

In this work we present an analytical model allowing us to predict the mass of the peptide cluster centres. The parameters of the model include: the frequencies of the amino acids in the sequence database (54), the average protein length of

the proteins in the database, the cleavage sites of the proteolytic enzyme and the cleavage probability. Based on this model we introduced a measure of deviation of peptide masses from the nearest cluster centre, which is a refinement of a measure proposed by Wool and Smilansky (30). Using this distance measure, we developed a calibration procedure which employs least squares linear regression in order to determine the affine model of the mass measurement error and subsequently to calibrate the spectra. Using this method we reached higher calibration accuracy as reported by Wool and Smilansky (30), and Gras et al. (35). We used the same distance measure to identify and remove non-peptide peaks prior to database searches performed by the Mascot search engine (55).

3.2 Methods

3.2.1 Data sets

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

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

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

3. A plant proteome of *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 aforementioned 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 a delayed ion extraction and by employing the MALDI AnchorChip ™targets (Bruker Daltonics, Bremen, Germany). Positively charged ions in the m/z range of $700 - 4,500 m/z$ 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.

3.2.2 Calibration

In order to perform filtering of non-peptide peaks the dataset must be calibrated to high mass measurement accuracy. To align the dataset we used a calibration sequence (41) consisting of several calibration procedures.

First calibration using external calibration samples was performed in order to remove higher order terms of the mass measurement error (40). Next, the affine mass measurement error of all samples on the sample support was determined by linear regression on the peptide mass rule introduced here. Subsequently, the thin plate splines were used to model the mass measurement error in dependence of the sample support positions to calibrate the spectra. Finally, the spectra were aligned using a modified spanning tree algorithm (41).

Mascot Database Search

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 $\pm 0.1 Da$. 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.

3.2.3 Sequence databases

We determined the amino acid frequencies of the nine protein sequence databases listed in Table 3.2. Seven of these databases are organism specific subsets of the $NCBI$ non-redundant protein database (107).

3.2.4 In Silico Protein Digestion

The theoretical digestion of the protein databases was done with ProtDigest (110), a command line program taking a protein sequence database file in *fasta* format and cleavage specificities as input. Other optional input parameters included fixed

Organizm	*length*	f_F	f_S	f_T	f_N	f_K	f_Y	f_E	f_V	f_Q	f_M
Arabidopsis t.	422.40	4.27	9.01	5.11	4.41	6.36	2.86	6.74	6.69	3.52	2.44
Drosophila m.	506.20	3.48	8.33	5.68	4.80	5.70	2.91	6.41	5.88	5.21	2.33
Escherichia coli	300.30	3.86	6.25	5.67	4.26	4.59	2.96	5.65	6.91	4.40	2.67
Homo sapiens	360.40	3.61	8.61	5.55	3.55	5.54	2.86	6.81	6.02	4.80	2.12
Mus musculus	378.30	3.74	8.58	5.55	3.59	5.71	2.88	6.75	6.11	4.74	2.22
Rattus norvegicus	484.40	3.81	8.33	5.52	3.59	5.62	2.74	6.77	6.32	4.64	2.28
Saccharomyces c.	447.00	4.47	9.02	5.93	6.18	7.26	3.41	6.43	5.58	3.94	2.10
Rhodopirellula b.	314.70	3.70	7.37	5.85	3.37	3.44	2.09	6.02	7.05	4.04	2.43
SwissProt DB	367.90	4.03	6.89	5.47	4.22	5.93	3.09	6.59	6.70	3.93	2.38
Mean	397.96	3.89	8.04	5.59	4.22	5.57	2.87	6.46	6.36	4.36	2.33
SD	71.90	0.32	0.98	0.24	0.88	1.07	0.35	0.39	0.50	0.54	0.18
Min	300.30	3.48	6.25	5.11	3.37	3.44	2.09	5.65	5.58	3.52	2.10
Max	506.20	4.47	9.02	5.93	6.18	7.26	3.41	6.81	7.05	5.21	2.67
	reference	f_C	f_L	f_A	f_W	f_P	f_H	f_D	f_R	f_I	f_G
Arabidopsis t.	(107)	1.80	9.52	6.36	1.26	4.80	2.28	5.43	5.39	5.34	6.41
Drosophila m.	(107)	1.95	9.02	7.36	1.00	5.46	2.64	5.18	5.53	4.96	6.17
Escherichia coli	(107)	1.17	10.23	9.27	1.50	4.32	2.22	5.21	5.54	5.94	7.38
Homo sapiens	(107)	2.24	9.78	6.98	1.35	6.22	2.51	4.73	5.64	4.28	6.80
Mus musculus	(107)	2.29	9.92	6.86	1.29	6.03	2.57	4.76	5.51	4.38	6.54
Rattus norvegicus	(107)	2.29	10.07	6.88	1.25	5.97	2.58	4.77	5.59	4.51	6.49
Saccharomyces c.	(107)	1.30	9.52	5.51	1.04	4.39	2.18	5.76	4.41	6.58	5.00
Rhodopirellula b.	(108)	1.27	9.31	9.25	1.54	5.33	2.31	6.23	6.96	4.95	7.48
SwissProt	(109)	1.57	9.63	7.80	1.17	4.86	2.27	5.30	5.29	5.92	6.94
Mean		1.76	9.67	7.36	1.27	5.26	2.40	5.26	5.54	5.21	6.58
SD		0.45	0.38	1.25	0.18	0.71	0.18	0.50	0.65	0.80	0.74
Min		1.17	9.02	5.51	1.00	4.32	2.18	4.73	4.41	4.28	5.00
Max		2.29	10.23	9.27	1.54	6.22	2.64	6.23	6.96	6.58	7.48

Table 3.2: Protein lengths and amino acid frequencies (one letter code) for nine in the nine databases. *length* – average protein length in database, $reference$ – database reference; f_i – amino acid frequencies

as well as variable modifications and number of missed cleavages. The output file contains all theoretically resulting peptides with their corresponding masses.

3.2.5 Regression analysis

The complete tryptic *insilico* digest of the SwissProt (109) database generated more than 7 million peptides. In order to compute the slope coefficient we were sampling 500 times 10000 monoisotopic and corresponding nominal masses. For each sample we fitted the affine linear model with and without fixed intercept using linear regression. The slope and intercept coefficients in Figure 3.1 are the

medians of these 500 samples.

3.2.6 Wool and Smilanskys algorithm

Wool and Smilansky (30) use a DFT to determine the calibration coefficients. The wavelength λ of a peptide peak-list can be determined by convolution. The "time domain" is the peak-list X with masses x_i. We computed the amplitude A (Equation 3.5) for a small range of frequencies $\omega \sim f = 1/\lambda$ around λ_{theo}. We scanned the range $\lambda \in \lambda_{theo} \pm 0.0005$ in steps of $5 \cdot 10^{-7}$ computing, for each λ, the real part (Equation 3.4), the imaginary part (Equation 3.3) and the amplitude $A(\omega)$(Equation 3.5):

$$f = 1/\lambda \quad \omega = 2\pi f \ , \tag{3.2}$$

$$\Im(\omega) = \sum_i sin(\omega x_i) \ , \tag{3.3}$$

$$\Re(\omega) = \sum_i cos(\omega x_i) \ , \tag{3.4}$$

$$A(\omega) = \sqrt{\Im(\omega)^2 + \Re(\omega)^2} \ . \tag{3.5}$$

The wavelength of the masses in the peak-list is the λ at the maximum of $A(\omega)$. The phase for this $\omega_0 = \omega_{\max A(\omega)}$ can be determined by:

$$\varphi_0 = \varphi(\omega_{\max A(\omega)}) = \arctan(\frac{\Im(\omega_0)^2}{\Re(\omega_0)^2}) \ . \tag{3.6}$$

The peak centres are at the line:

$$\acute{M} = \frac{2 \cdot \pi}{\omega_0} \cdot N + \frac{\varphi_0}{\omega_0} \quad \text{where} \quad N = 1, 2, \ldots, n \ . \tag{3.7}$$

But they should be on the line:

$$M = \lambda_{theo} * N \ . \tag{3.8}$$

Solving Equation 3.7 for N and substituting N in the Equation 3.8 yields the Equation:

$$M = \frac{\lambda_{theo} \cdot \omega_0}{2 \cdot \pi}(\acute{M} - \frac{\varphi_0}{\omega_0}), \tag{3.9}$$

$$\alpha = \frac{\lambda_{theo} \cdot \omega_0}{2 \cdot \pi} \quad \text{and} \quad \beta = \frac{\varphi_0}{\omega_0} \text{ and} \tag{3.10}$$

$$m_{corr} = \alpha(m_{exp} - \beta) = \alpha m_{exp} - \alpha\beta, \tag{3.11}$$

which can be used to correct the masses. This is an affine linear model with two coefficients α and $\alpha\beta$.

3.3 Results and Discussion

3.3.1 A simple way to predict the peptide mass cluster centres of a protein database

Figure 3.1 shows the mass defect, the difference of the monoisotopic ($m^{(M)}$) and nominal ($m^{(N)}$) masses of peptides of a sequence specific *in silico* protein sequence database digest (109), as a function of $m^{(N)}$. The peptides were produced with the restriction that no missed cleavages were allowed. A strong linear dependence of the mass defect on $m^{(N)}$ can be observed.

The first model of this dependence which we examined was $m^{(M)} - m^{(N)} = c_1 \cdot m^{(N)}$. We fixed the intercept at 0, because a hypothetical peptide with a nominal mass of 0 must have a monoisotopic mass equal to 0. The slope coefficient c_1, determined by linear regression (cf. Methods) equalled $4.98 \cdot 10^{-4}$(Figure 3.1, Panel A – red dashed line), which is a value similar to the values $4.95 \cdot 10^{-4}$ reported by Wool and Smilansky (30).

We were interested in determining the dependence between monoisotopic and nominal mass analytically. For example, the monoisotopic mass ($m^{(M)}$) of hypothetical peptides built only of one amino acid i can be predicted, given their nominal mass ($m^{(N)}$) by $m_i^{(M)} = \lambda_i m_i^{(N)}$ when $\lambda_i = m_i^{(M)}/m_i^{(N)}$. For peptides generated by random cleavage of protein sequences from a protein database this dependence is approximated by:

$$\lambda_{DB} = \frac{\sum_{i \in AA} f_i m_i^{(M)}}{\sum_{i \in AA} f_i m_i^{(N)}}\ , \tag{3.12}$$

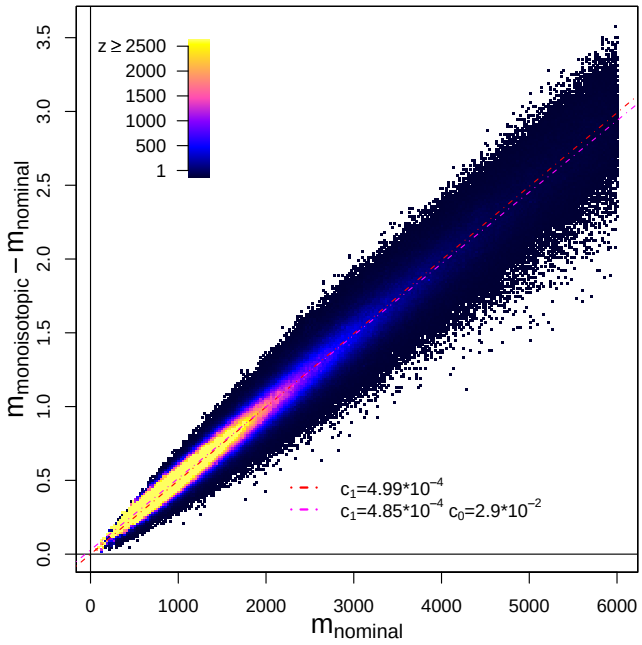

Figure 3.1: The peptide mass rule. Panel A: Scatterplot of $m^{(M)} - m^{(N)}$ against the $m^{(N)}$ mass ($m^{(M)}$ - monoisotopic mass, $m^{(N)} - nominalmass$). Inset top left - colour coded number z of peptide masses per 0.25 pixel. Red dashed line – the model determined by linear regression with intercept fixed at 0. The magenta line represents the cluster centres predicted by linear regression.

where f_i is the frequency of the amino acid i in the database.

Now write $m_i^{(M)} = \lambda_{DB} m_i^{(N)} + \epsilon_i$. Substituting this is (3.12), it follows that $\sum_{i \in AA} f_i \epsilon_i = 0$. Therefore, for an amino acid randomly selected from the database, with frequencies f_i, the expectation of ϵ_i is zero. Now consider a peptide made of a random selection of J amino acids, $i(1), \ldots, i(J)$. The ratio of monoisotopic to nominal mass for this peptide would be:

$$\lambda_p = \frac{\sum_{j=1}^{J} m_{i(j)}^{(M)}}{\sum_{j=1}^{J} m_{i(j)}^{(N)}} = \frac{\lambda_{DB} \sum_{j=1}^{J} m_{i(j)}^{(N)} + \sum_{j=1}^{J} \epsilon_{i(j)}}{\sum_{j=1}^{J} m_{i(j)}^{(N)}}.$$

If $\sum_i \epsilon_{i(j)}$ were uncorrelated with $(\sum_i m_{i(j)}^{(N)})^{-1}$ for a random selection of amino acids, then λ_p would have expectation λ_{DB}. Of course, there may be a relationship between ϵ_i and $m_i^{(N)}$ and we would wish to use any such relationship to improve

prediction of $m_i^{(M)}$.

Figure 3.2 visualises the frequencies f_i of all amino acids in the *Uniprot* database (109) with their respective λ_i plotted on the abscissa. The position of the red vertical line on the abscissa denotes λ_{DB} (Equation 3.12) and equals $\lambda_{DB} = 1.000511$. The dotted, dashed and dot dashed lines indicate the wavelength λ of DHB, alpha-cyano and sinapic acid mass spectrometric matrix clusters, respectively.

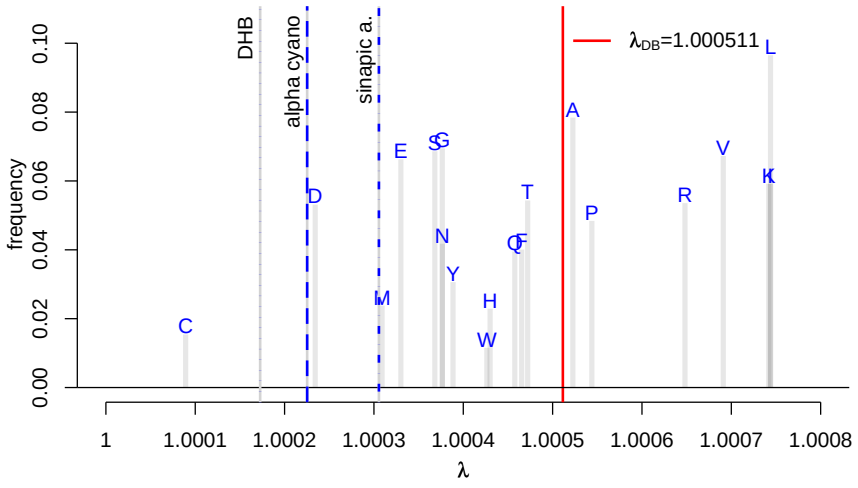

Figure 3.2: Bar-plot of the Amino Acid frequencies. The bars are drawn on the position of $\lambda_i = m_i^{(M)}/m_i^{(N)}$, for each amino acid i. The red line indicates λ_{DB} computed using the Equation 3.12. Dotted blue line – λ_{DHB} 2,5-dihydroxybenzoic acid; dashed line – $\lambda_{alphacyano}$ alpha-Cyano-4-hydroxycinnamic acid; dot dashed line – $\lambda_{sinapica.}$ 3,5-Dimethoxy-4-hydroxycinnamic acid.

When testing for the significance of the intercept coefficient in the regression model $m_M \propto \lambda m_N$ of a sequence specific (Tryptic) *in silico* database digest, we found that the intercept coefficient must be included into the model. Therefore, the extended model of the monoisotopic peptide mass cluster centres was:

$$m^{(M)} = c_1 \cdot m^{(N)} + c_0 \ . \tag{3.13}$$

Subtracting m_N from each side of Equation 3.13 we obtained $\Delta = m^{(M)} - m^{(N)} = (c_1 - 1) \cdot m^{(N)} + c_0$. The coefficients of the affine linear model of the cluster

centres, determined using regression analysis of $\Delta = m^{(M)} - m^{(N)}$ on $m^{(N)}$ were $c_0 = 0.029$ and $(c_1 - 1) = 4.85 \cdot 10^{-4}$.

The maximal difference between the prediction of $m^{(M)}$ using $m^{(M)} = 1.000499 \cdot m^{(N)}$ and $m^{(M)} = 1.000485 \cdot m^{(N)} + 0.029$ is 0.022 Dalton for $m^{(N)} \in [600, 2500]$ Dalton.

The influence of the digestion enzyme on the wavelength of peptide mass clusters

In case of a complete sequence specific cleavage of proteins, the number of generated peptides is C_P+1 peptides, given that C_P is the number of cleavage sites per protein. The peptides generated from the terminus of the protein (further called *terminal*) will not bear a cleavage site residue R_C at their end. All the other peptides, which we call *internal*, will have such a residue at their end. The fraction of the internal peptides $f_{c,n}$ is given by

$$f_{c,n} = \frac{C_P - n}{C_P + 1 - n} \ , \tag{3.14}$$

where n is the number of missed cleavages per protein. We approximate C_P, for a sequence database, by:

$$C_P = |P| \cdot (\sum f_{R_C}), \tag{3.15}$$

where f_{R_C} are the relative frequencies of the cleavage sites and $|P|$ is the average protein length in the database. The fraction of the terminal peptides in case of n missed cleavages is given by $1 - f_{c,n}$.

The fraction of cleavage site residues R_C in a internal peptide of mass m_{pep}, with n missed cleavage sites is denoted $f_{m,n}$ and approximated by:

$$f_{m,n} = (n+1)\frac{\bar{m}}{m_{\text{pep}}} \ , \tag{3.16}$$

where $\bar{m}$ is the average mass of an amino acid residue. A more accurate model of $f_{m,n}$ is provided in the *Appendix*. In the case of terminal peptides the fraction of cleavage site residues R_C equals $f_{m,n-1}$. The fraction of all the other amino acid

residues $R \backslash R_C$ equals $1 - f_{m,n}$ or $1 - f_{m,n-1}$ respectively. Table 3.3 summarises these results.

$R_{\text{non-cleavage}}$	R_{cleavage}		Peptide type
$(1 - f_{m,n})$	$f_{m,n}$	$f_{c,n}$	internal
$(1 - f_{m,n-1})$	$f_{m,n-1}$	$1 - f_{c,n}$	terminal

Table 3.3: Frequencies of cleavage site residues, and all other residues, in peptides of mass m and of terminal, and internal, peptides. R_{cleavage} – frequencies of cleavage site residues; $R_{\text{non-cleavage}}$ – frequencies of non-cleavage site residues; $f_{m,n}$ – see Equation 3.16; $f_{c,n}$ – see Equation 3.14.

In the case of internal peptides, the average contribution of the amino acid residues to the peptide mass is the weighted sum:

$$m^{(*)}_{R_C,n} = (1 - f_{m,n}) \cdot m_{none} + f_{m,n} \cdot m_{R_c} \tag{3.17}$$

$$= m_{none} + f_{m,n} \cdot (m_{R_C} - m_{none}) \ , \tag{3.18}$$

where

$$m_{none} = \sum_{i \in R \backslash R_C} f_i \cdot m_i, \tag{3.19}$$

is the average mass of non cleavage residues, and:

$$m_{R_C} = \sum_{i \in R_C} f_i \cdot m_i. \tag{3.20}$$

is the average mass of the cleavage site residues R_C. Finally, the wavelength of internal peptides is presented as:

$$\lambda^{m}_{R_C,n} = \frac{m^{(M)}_{R_C,n}}{m^{(N)}_{R_C,n}} \tag{3.21}$$

The wavelength of terminal peptides was determined by: $\lambda^{(n-1)}_{R_C,m} = \frac{m^{(M)}_{R_C,n-1}}{m^{(N)}_{R_C,n-1}}$.

The wavelength λ of all peptides at a mass m with exactly n missed cleavages is given by:

$$\lambda_{R_C,n}^{m,*} = \frac{m_{R_C,n}^{(M),*}}{m_{R_C,n}^{(N),*}} \tag{3.22}$$

where

$$\begin{aligned}
m_{R_C,n}^{[MN],*} &= f_{c,n} \cdot m_{R_C,n}^{[MN]} + (1 - f_{c,n}) \cdot m_{R_C,n-1}^{[MN]} && (3.23)\\
&= m_{none} + (m_{R_C} - m_{none}) \cdot (f_{c,n} f_{m,n} + f_{m,(n-1)} - f_{c,n} f_{m,(n-1)}) && (3.24)\\
&\underbrace{=}_{\text{with Eq. 3.16}} m_{none} + \frac{\bar{m}}{m}(f_{c,n} + n)(m_{R_C} - m_{none}) && (3.25)\\
&\underbrace{=}_{\text{with Eq. 3.14}} \boldsymbol{m_{none} + \left(\frac{C_p - n}{C_p + 1 - n} + n\right) \cdot \frac{\bar{m}}{m} \cdot (m_{R_C} - m_{none})} && (3.26)
\end{aligned}$$

is the weighted sum of the mass of the terminal peptides (with frequency $1 - f_{c,n}$) and the internal peptides (with frequency $f_{c,n}$).

Cleavage probability p_c In practice, the cleavage probability will depend on various factors, for example on the incubation time and the efficiency of the protease used. The probability to generate a peptide with $n \in 0...\infty$ missed cleavage sites, given the cleavage probability p_c can be modelled using the geometric distribution:

$$P(n, p_c) = (1 - p_c)^n \cdot p_c \tag{3.27}$$

Furthermore,

$$\sum_{n=0}^{\infty} (1 - p_c)^n \cdot p_c = 1 \tag{3.28}$$

holds. Hence, given the cleavage probability is p_c and cleavage residues R_C, we express the peptide mass by:

$$m_{R_C,p_c}^{*} = \boldsymbol{m_{none} + \sum_{n=0}^{\infty} (1 - p_c)^n \cdot p_c \cdot (m_{R_C} - m_{none}) \cdot S_n} \ , \tag{3.29}$$

where

$$S_n = (f_{c,n} f_{m,n} + f_{m,(n-1)} - f_{c,n} f_{m,(n-1)}). \tag{3.30}$$

Therefore, the wavelength λ of peptides if the cleavage probability is p_c is given by:

$$\lambda^{m,*}_{R_C,p_c} = \frac{m^{(M),*}_{R_C,p_c}}{m^{(N),*}_{R_C,p_c}} \tag{3.31}$$

The monoisotopic mass as a function of the nominal mass can be expressed by:

$$m^{(M)} = \lambda^{(m),*}_{R_C,p_c} \cdot m^{(N)} \tag{3.32}$$

$$= \frac{m^{(M),*}_{R_C,p_c} \cdot m^{(N)}}{m^{(N),*}_{R_C,p_c}} \tag{3.33}$$

$$\underbrace{=}_{\text{Eq. 3.30, 3.14}} \frac{m^{(M)}_{none} \cdot m^{(N)} + \sum_{n=0}^{\infty}(1-p_c)^n \cdot p_c \cdot (m^{(M)}_{R_C} - m^{(M)}_{none}) \cdot \bar{m}(f_{c,n}+n)}{m^{(N)}_{none} + \sum_{n=0}^{\infty}(1-p_c)^n \cdot p_c \cdot (m^{(N)}_{R_C} - m^{(N)}_{none}) \cdot \frac{\bar{m}}{m^{(N)}}(f_{c,n}+n)} \tag{3.34}$$

$$\underbrace{\approx}_{\text{for } m^{(N)} \gg \bar{m}} \frac{m^{(M)}_{none} \cdot m^{(N)}}{m^{(N)}_{none}} + \frac{\sum_{n=0}^{\infty}(1-p_c)^n \cdot p_c \cdot (m^{(M)}_{R_C} - m^{(M)}_{none}) \cdot \bar{m}(f_{c,n}+n)}{m^{(N)}_{none}} \tag{3.35}$$

This equation represents our final model of the peptide mass cluster centres. To illustrate the accuracy of the prediction we computed the residuals Δ between the monoisotopic masses of the *in silico* database digest and the cluster centres predicted by Equation 3.34. Figure 3.3 shows the relative residuals $\Delta^{ppm}(m) = \Delta(m)/m \cdot 10^6$, in parts per million. The grey line shows the moving average of the residuals $\Delta^{ppm}(m)$ computed for a window of $15Da$.

Figure 3.4, panel A, shows the difference between nominal and monoisotopic mass $(m^{(M)} - m^{(N)})$ where $m^{(M)}$ was predicted using the model of Equation 3.34. We observed that $m^{(M)} - m^{(N)} \propto m^{(N)}$ is approximately a straight line for the mass range greater than $500Da$. By using the predicted monoisotopic mass $m^{(M)}$ at $m^{(N)} = 500$ and at $m^{(N)} = 3000$ we determined the slope:

$$c_1 = \frac{3000 \cdot \lambda^{(3000),*}_{R_C,p_c} - 500 \cdot \lambda^{(500),*}_{R_C,p_c}}{3000 - 500} = 1.000482, \tag{3.36}$$

and intercept coefficient

$$c_0 = 500 \cdot (\lambda^{(500),*}_{R_C,p_c} - 1) - c_1 \cdot 500 = 0.029. \tag{3.37}$$

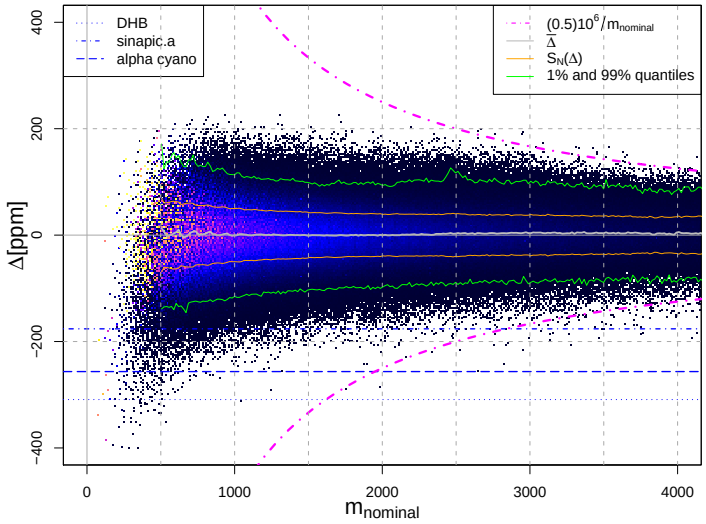

Figure 3.3: Deviation Δ^{ppm} of peptide masses from mass cluster centres predicted using the Equation 3.34 in parts per million [ppm]. Gray line – moving average of Δ^{ppm}. Orange lines – Standard deviation of Δ^{ppm}, Green lines – 1% and 99% Quantile computed for mass windows having a size of $15Da$ and covering the mass range. Magenta dot dashed line – maximum possible deviation from cluster centre, which can be assigned to the true cluster centre using the Equation 3.40. Horizontal dotted blue line – distance of DHB (2,5-dihydroxybenzoic acid) matrix clusters from the peptide mass cluster centres; dashed line – distance of *alphacyano* (alpha-Cyano-4-hydroxycinnamic acid) clusters from the peptide mass cluster centres; distance of *sinapicacid* (3,5-Dimethoxy-4-hydroxycinnamic acid) clusters from peptide mass cluster centres.

These coefficients are in good agreement with the slope and intercept determined by linear regression for the *in silico* sequence database digest (Figure 3.1).

Furthermore, we observed that the intercept c_0 will be positive if $m_{R_C} > m_{none}$, zero or negative otherwise. The slope c_1 equals $\lambda_{none} = \frac{m_{none}^{(M)}}{m_{none}^{(N)}}$, for large $m^{(N)}$, because the frequency of the cleavage site residues R_C decreases with increasing peptide length:

$$\lim_{|Pep|\to\infty} f_{m,n} \propto \lim_{m_{pep}\to\infty} \frac{(n+1)\bar{m}}{m^{(N)}} = 0 \ .$$

Figure 3.4, panel B, displays the difference between the line $(c_1+1)\cdot m^{(M)}+c_0$ and the prediction made using Equation 3.13. For the mass range $m \in (500, 4000)$ where peptide masses for peptide mass fingerprinting are acquired this difference is minimal.

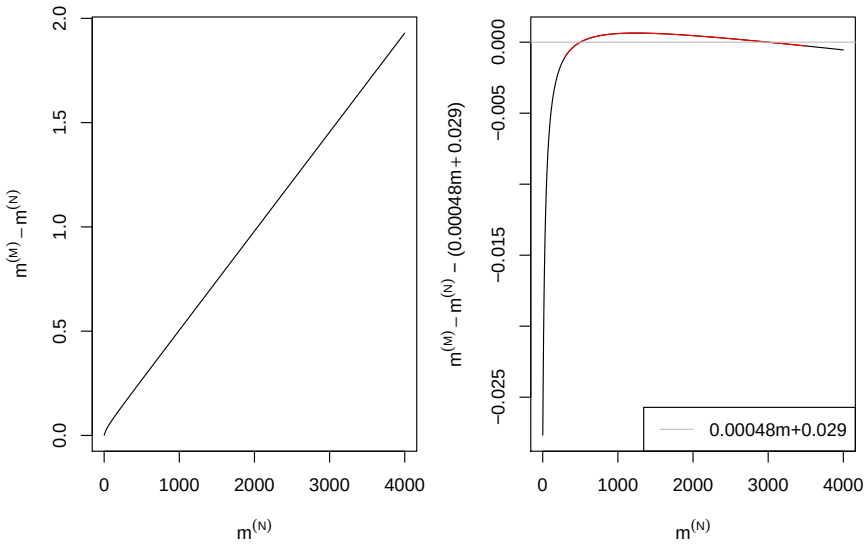

Figure 3.4: The monoisotopic mass as an function of the nominal mass. Left panel : $m^{(M)} - m^{(N)} = (\lambda^{(m),*}_{R_C,p_c} - 1)\cdot m^{(N)}$. Right panel : Difference between $(\lambda^{(m),*}_{R_C,p_c} - 1)\cdot m^{(N)}$ and $0.00048 \cdot m^{(N)} + 0.029$.

The coefficients c_0 and c_1 do not depend on the mass of the peptides. Due to this feature, we are going to use the affine model $c_1 m^{(N)} + c_0$ to predict the peptide mass cluster centres in the applications discussed later. This simplified model is also in agreement with the affine model (Equation 3.13), which has been fitted by linear regression to the *in silico* database digest in order to explain the dependency of the peptide mass cluster centres on the nominal mass.

3.3.2 Error of the model

Combinatorial restrictions may cause significant differences between the linear prediction of the model (Equation 3.34) introduced and the actual location of the cluster centre. To assess this error we first computed the location of the cluster centres (average of all monoisotopic masses in cluster) of the *in silico* database

digest, and afterwards determined the difference to the cluster centre location predicted by model of Equation 3.34. This difference $\bar{\Delta}(cluster)$ is shown in Figure 3.5.

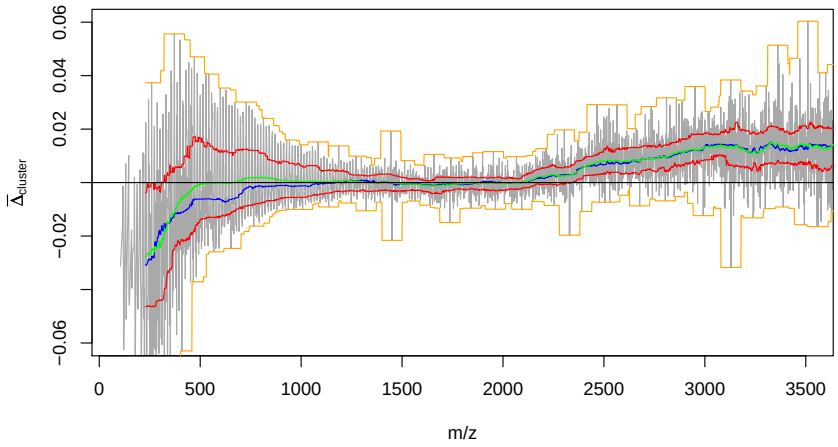

Figure 3.5: Difference between cluster centre computed for the *in silico* database digest and the cluster centre location predicted by the model (Equation 3.34). Orange lines – minimum and maximum, red lines – first and third quartile, green – mean, blue – median of the differences computed for a moving window of $100Da$.

For a moving window of $100Da$ we computed the maximum and minimum (orange), third and first quartile (red), median (blue) and mean(gree) of $\bar{\Delta}(cluster)$. The combinatorial restriction decreases with increasing mass and for peptide masses greater than $1000Da$ it is negligible. However, $\bar{\Delta}(cluster)$ increases again for masses greater than $2500Da$ because peptide masses may deviate more strongly from the cluster centres and furthermore much fewer long peptides are generated.

3.3.3 The type of distribution around the cluster centres

In order to remove non-peptide peaks prior to database search, filtering thresholds have to be chosen. In Figure 3.3 the orange line visualises the standard deviation while the green lines show the 1% and 99% quantiles of $\Delta^{ppm}(m) = \Delta(m)/m \cdot 10^6$

computed for a mass window of $15Da$. In addition the dotted, dashed, and dot dashed line show the deviation $\Delta^{ppm}(m)$, at which clusters of mass spectrometric matrices are expected.

The standard deviation of $\Delta^{ppm}(m)$ is symmetric and does not change for $m > 1500$. We were interested to determine the distribution of Δ^{ppm} around the peptide mass cluster centres. To determine the type of distribution we use qqplots (111) shown in Figure 3.6. We compared the distribution of the residues $\Delta^{ppm}(m)$, observed for four different mass windows ($m \in (500 - 530)$, $m \in (1000 - 1110)$, $m \in (2000 - 2200)$ and $m \in (3400 - 3700)$) with the normal distribution and t-distributions with various degrees of freedom. The t-distribution with degrees of freedom $\mu \in (15, 25)$ is a good approximation of the empirical distribution of Δ^{ppm} for masses > 2000,.

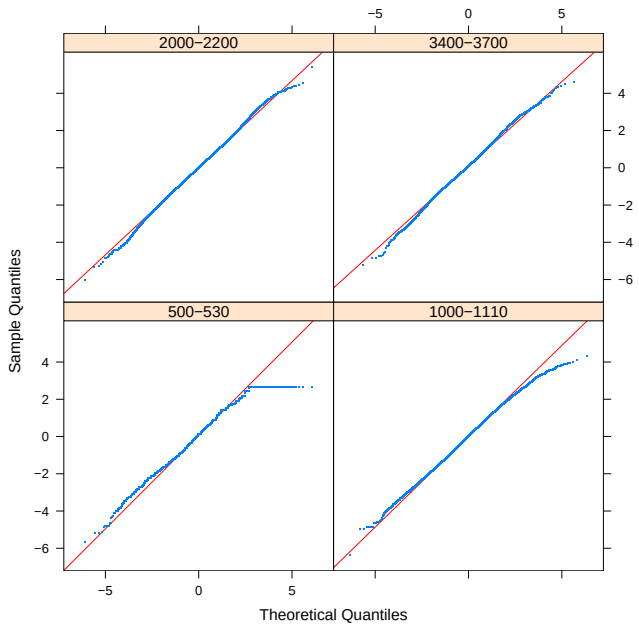

Figure 3.6: qqplot - of $\Delta^{ppm} = m_M - c_1 \cdot m_N - c_0$ versus the t-distribution with 19 degrees of freedom for four mass ranges $m \in (500 - 530), m \in (1000 - 1110), m \in (2000 - 2200) \text{and} m \in (3400 - 3700)$.

3.3.4 Sensitivity analysis

The input parameters to the model of the peptide mass cluster centres included:

- f_i – frequencies of the amino acids.
- cleavage specificity of the protease R_C
- $|P|$ - Protein length
- p_c - cleavage probability

To examine how the output of the model is influenced by these factors we varied the protein length $|P|$ in steps of 100 from 300 to 800 amino acids per protein. We determined the amino acid frequencies f_i for 9 sequence databases (cf. Methods) and used them as inputs to the model. Furthermore, six cleavage specificities (shown in Table 3.4) were examined and the cleavage probability p_c was changed from 0.4 to 1 in increments of 0.2.

	Enzyme	R_C
1	Trypsin/P	K,R/P
2	Arg.C	R/P
3	CNBR + Trypsin	F,Y,M
4	Lys-C	K/P
5	PepsinA	F,L
6	CNBr	M

Table 3.4: Cleavage sites of proteolytic enzymes (4)

The box-plots, of Figure 3.7, Panel A demonstrate that the values of the intercept coefficient c_0 (Equation 3.37) mainly depend on the cleavage probability p_c and on the cleavage specificity of the proteolytic enzyme. The relatively small height of the boxes indicates that the differences in amino acid frequencies f_i for the databases examined, and the average protein length $|P|$ have a negligible effect on the intercept coefficient. The slope coefficient c_1 (see Equation 3.36) depends only on the cleavage site specificities of the proteolytic enzyme and the amino acid frequencies f. The box-plots 3.7 Panel B show that the model output is highly sensitive to the cleavage specificity of the proteolytic enzyme.

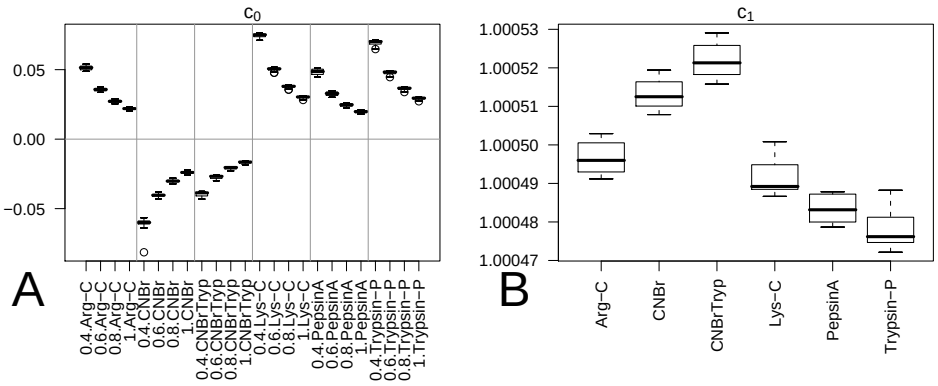

Figure 3.7: Panel A – Box plots of the intercept coefficient c_0 (Equation 3.37) itemised according the cleavage specificity and cleavage probability. Panel B – Box plots of the slope coefficient c_1 (Equation 3.36) itemised according the cleavage specificity.

3.3.5 A measure of distance to cluster centres

Given an experimentally determined m_M we were interested to estimate the deviation Δ from the closest predicted cluster centre. The model of the monoisotopic mass is:

$$c_0 + c_1 \cdot m_N + \Delta = m_M, \tag{3.38}$$

where c_0, c_1 can be obtained using the Equations 3.37 and 3.36, m_N is the nominal mass (an integer).

Therefore, for a given m_M, c_0 and c_1 we can determine the deviation Δ from the closest cluster centre of *smaller* mass by using the modulo operator as suggested by Wool and Smilansky (30):

$$(m_M - c_0)(\mathrm{mod} c_1) = (c_1 \cdot m + \Delta)(\mathrm{mod} c_1) = \Delta\ . \tag{3.39}$$

However, in order to determine the distance to the closest cluster centre we considered two cases:

$$\Delta_\lambda(m_i, 0) = \begin{cases} (m_i - c_0)(\mathrm{mod} c_1) & \text{if} \quad (m_i - c_0)(\mathrm{mod} c_1) < 0.5 \\ -1 + (m_i - c_0)(\mathrm{mod} c_1) & \text{otherwise}\ . \end{cases} \tag{3.40}$$

The units of $\Delta_\lambda(m_i, 0)$ are in $[m/z]$. The magenta dot dashed curves in Figure 3.3 indicate the maximum detectable distance from cluster centres in ppm ($\pm 0.5Da/m \cdot 10^6[ppm]$). Deviations from the cluster centres outside the range enclosed by these two curves are assigned to the wrong cluster. In case of theoretical peptide masses and experimental masses calibrated to high precision, such distances are observed only for masses greater than $2500Da$. Fortunately, the majority of tryptic peptide masses detected in a mass spectrometric peptide fingerprint experiment are below this mass.

3.3.6 Applications

3.3.6.1 Linear Regression on Peptide Mass Rule LR/PR

The limitations of calibration methods based on the property of peptide mass clustering are a mass accuracy of only $0.2Da$, its sensitivity to non-peptide peaks in the spectra, and that it completely fails if the number of peptide peaks in the peak list is small (30; 35; 43). Hence, in practice, the method is used to confirm the results of internal calibration only (43; 112). However, the advantage of the calibration methods based on the property of peptide mass clustering, over other calibration methods (41), is that no internal or external calibrants are required in order to calibrate the peptide mass lists.

We propose here a novel method for the calibration of PMF data, based on robust linear regression and the distance measure introduced in the Equation 3.40. To determine the slope of the mass measurement error we computed the deviation from the peptide mass rule for every pair of peak masses (m_i, m_j) within a peak-list, employing the following equation:

$$\Delta_\lambda(m_i, m_j) = \begin{cases} |m_i - m_j|(\mathrm{mod} c_1) & \text{if} \quad |m_i - m_j|(\mathrm{mod} c_1) < 0.5 \\ -1 + (|m_i - m_j|(\mathrm{mod} c_1)) & \text{otherwise} \quad , \end{cases} \tag{3.41}$$

with c_1 given by Equation 3.36.

Figure 3.8 left top panel shows the distance $\Delta_\lambda(m_i, m_j)$ (Equation 3.41) as a function of $\Delta_d = |m_i - m_j|$, computed for all pairs $(m_i, m_j) \in$ peak-list, which adhere to the additional constraint that $\Delta_d = |m_i - m_j| < m_{max}$. This constraint

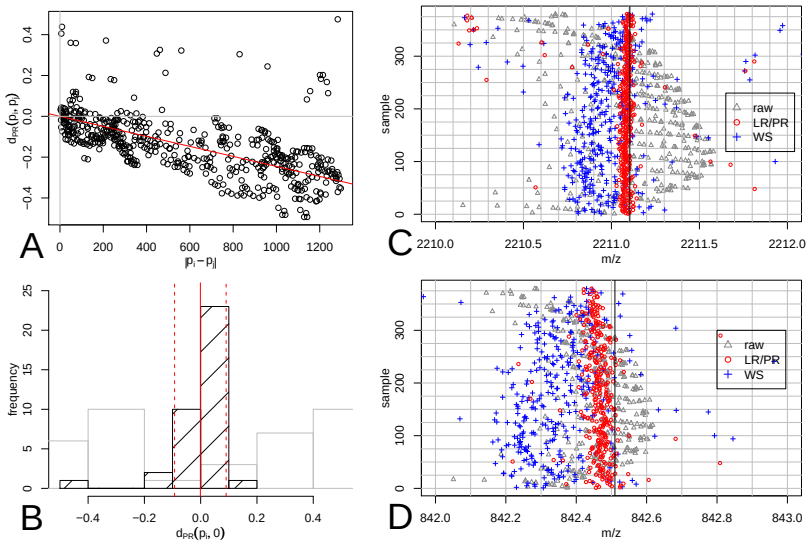

Figure 3.8: Principle and results of linear regression on peptide rule *LR/PR* calibration. Panel A: Scatter-plot of $\Delta_{PR}(m_i, m_j)$ (Equation 3.41) in dependence of $\Delta_d = |m_i - m_j|$. The slope, obtained by robust regression, is shown by the red line. Panel B: Histogram (black with diagonals) of $d_{PR}(m_i, 0)$. The continuous vertical red line denotes the average ($\bar{d}_{PR}(m_i, 0)$) and the dotted vertical lines denote $\bar{d}_{PR}(m_i, 0) \pm S_N$. The histogram in gray is showing the distribution of $d_{PR}(m_i, 0)$ previous to removing the slope error (see text). Panel C & D: Strip-charts of the data-set for a mass range of $2210 - 2212Da$ and $842 - 843Da$, including the tryptic autolysis peaks $842.508Da$ and $2211.100Da$. Gray triangles – raw data; blue "+" – Wool Smilansky algorithm (cf. Appendix); red "o" – LR/RP algorithm for tryptic peaks.

is necessary because the measure Δ_λ is only able to assign deviation smaller than $0.5Da$ to the correct cluster centre. For large values of Δ_d, Δ_λ increases, if $c_1 \neq 0$ and assignments to wrong clusters may occur. If a systematic dependence of $\Delta_\lambda$onΔ_d is observed it indicates a mass measurement error. We determined the slope $\hat{c}_1$ using robust linear regression (113) with the intercept fixed at 0. To correct the peak-list masses we applied

$$m_{corrected} = m_{experimental} \cdot (1 - \hat{c}_1) \ .$$

To determine the intercept coefficient of the mass measurement error we subsequently computed $\Delta_\lambda(m_{corrected}, 0)$ (using Equation 3.40), for all peak-list

masses. Figure 3.8 Panel B shows the distribution of $\Delta_\lambda(m_i, 0)$ before correcting for the slope error (gray histogram) and afterwards (black histogram). The red vertical line indicates the mean $\bar{\Delta}_\lambda(m_i, 0)$, computed for the corrected data, which we used to approximate the intercept $\hat{c}_0$ of the mass measurement error.

The strip charts (Figure 3.8, Panel C and D) visualises the experimental masses of two trypsin peptides $842.508Da$ and $2211.100Da$ observed in most of the samples of the dataset with 380 peak-lists. The result of LR/PR calibration (red circles) is compared with raw masses (gray triangles) and the output of the Wool and Smilansky calibration method (blue crosses). The LR/PR-method is able to calibrate mass spectrometric peak-lists to an accuracy of $0.1Da$. This measurement accuracy surpasses the other published calibration methods (30; 35) at least two-fold.

3.3.6.2 Filtering of non-peptide peaks using the peptide mass rule

Non-peptide peaks can be recognised according to their deviation from the cluster centres. The amino acids that have the most extreme λ values are I, L and K (because of their large fraction of Hydrogen H (1.007825) atoms) and C (Cysteine - because of the heavy sulfur atom S (31.97207)). If we plot the position after the decimal point given by $n \cdot (\lambda_i - 1) (\mathrm{mod} 1)$ with $n \in \mathbb{N}$, for $i = L$ and $i = C$, and connect the points for readability purposes by a line (the red and green lines in Figure 3.9 respectively), we obtain the range enclosing any possible decimal point a theoretical peptide mass can have. If a mass with a decimal point lying in the dashed region is detected it can not be a peptide peak. For peptide peaks, the following inequalities hold:

$$-413[ppm] = (\lambda_C - \lambda_{DB}) \cdot 10^6 \; < \; \Delta_\lambda(m, 0) \cdot 10^6 / m \; = \; \Delta_\lambda^{ppm}(m, 0) \; < \; (\lambda_L - \lambda_{DB}) = 241[ppm] \,, \tag{3.42}$$

where $\lambda_{DB} = 1.000511$ (Equation 3.12). We used the relative deviation of Δ^{ppm} from the cluster centre in parts per million instead of using absolute values.

Figure 3.3 shows that only very short peptides approach the lower bound of $-413ppm$. This is due to the low frequency of Cysteine (C). The high frequencies of K, L, I (whose $\lambda \approx 1.00074$) mean that the theoretical upper bound of $241ppm$ can indeed be reached by some peptides with a mass of $\approx 1000Da$. Peptides of higher mass never approach the upper and lower theoretical bound due to the rapidly decreasing probability to consist of $K, L or I$, or of C only. The lines for the

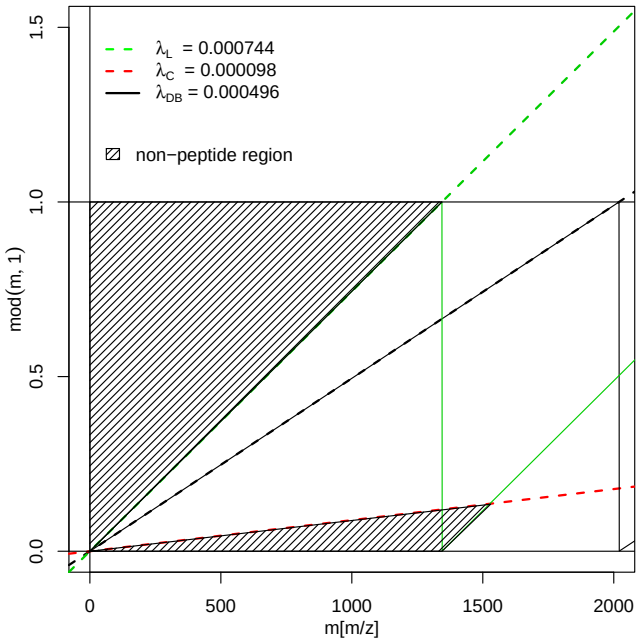

Figure 3.9: Schema of non-peptide mass filtering. Abscissae - peptide mass, ordinate – m mod 1, dashed region – non-peptide masses. Green line – decimal part of poly-(L(lys),I(ile)) peptide masses as a function of their mass. Red line – decimal part of poly-(C(cys)) peptide masses as function of their mass. Black line – Predicted cluster centres using the Equation 3.12.

standard deviation of S_N (orange lines) and of the 1% and 99% quantile (green lines) in Figure 3.3 indicate that it is an exceedingly rare event to encounter a peptide mass for which $\Delta_\lambda^{ppm}(m, 0)$ will deviate more than $200ppm$ from the peptide cluster centre predicted by our model. Therefore, we use $200ppm$ as a filtering threshold. An essential requirement, to apply this filtering method successfully is, that peak-list must be calibrated to high precision (41).

Figure 3.10 visualizes the result of non-peptide peak filtering in case of a dataset of 380 calibrated peak-lists. Spots removed by applying the filtering criterion $\Delta_\lambda^{ppm}(m, 0) > 200$ are shown in green. Peptide masses removed due to filtering of abundant masses (41) are shown in red.

We studied how the non-peptide peak filtering influences the *Probability Based Mascot Score* (PBMS) (55). In theory, for example one cystein rich

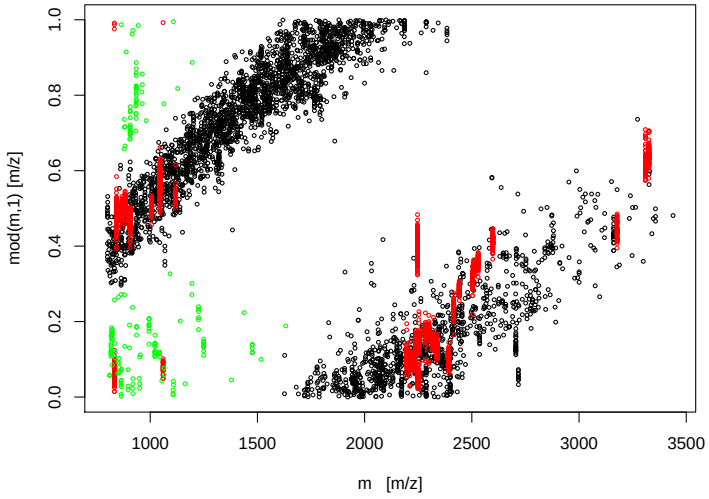

Figure 3.10: Scatter plot : abscissae - peptide mass m_i, ordinate - $m_i mod \lambda$ with $\lambda = 1.000495$. In red are highlighted peaks removed from the dataset because of their high frequencies. In green, peaks removed due to the strong deviation from the peptide mass cluster centres.

peptide strongly deviating from the peptide mass rule and with a unique mass in the database digest, if properly assigned is sufficient to identify the protein unambiguously (30). In case of PBMS, which requires multiple matches to peptide masses, a single match of a unique peptide mass, even if properly assigned, will not give a score indicating reliable identification of the protein. Furthermore, this scoring scheme takes into account the number of non-matching peaks. If many unassigned peaks are observed, the score is decreased and the assignment is interpreted as insignificant. Therefore, the removal of non-peptide peaks should increase the identification sensitivity.

Table 3.5 demonstrates that an increase of 2.5% in the number of identified samples can be obtained by removing all peaks with a distance $\Delta_\lambda^{ppm}(m,0) > 200ppm$ from the peptide peak-lists. Row 8 of Table 3.5 shows that non-peptide peak filtering increases the PBMS score in $30-55\%$ of cases. Removal of peptide peaks due to filtering caused a decrease of the PBMS score in less than 1% of samples.

We concluded that non-peptide peak filtering increases the sensitivity of protein identification if using the PBMS scoring schema. However, the extend

		Arabidopsis t.	*Rhodopirelulla b.*	*Mus musculus*
1	Identification no PR filtering	423	1009	872
2	Identification with PR filtering	432	1017	894
3	Change in identification (Percent)	2.13	0.79	2.52
4	Total nr. of samples*	818	1169	1709
5	Nr. samples with PBMS increase	240	622	724
6	Nr. samples with no change of PBMS	571	542	982
7	Nr. samples with PBMS decrease	7	5	3
8	Percent increase of PBMS score	29.34	53.21	42.36
9	Percent decrease of PBMS score	0.86	0.43	0.18

Table 3.5: Results for filtering of non-peptide masses. Columns: *Arabidopsis t.*, *Rhodopirelulla b.*, *Mus musculus* – peptide mass fingerprint datasets (cf. Methods). Row 1 – number of samples with a significant PBMS score prior to filtering of non-peptide peak masses. Row 2 – number of samples with a significant PBMS score for peak-lists with non-peptide removed. Row 3 – relative change of the identification rate (Row 2 − Row 1)/Row1 · 100. Row 4 – Total number of samples which produced a PBMS score. Row 5 – number of samples for which an increase of the PBMS score due to non peptide peak filtering was observed. Row 6 – number of samples for which no change of the PBMS score due to non-peptide peak filtering was observed. Row 7 – number of samples for which a decrease of the PBMS score due to non-peptide peak filtering was observed. Row 8-9 – relative increase and decrease of the PBMS score, respectively.

to which extend these results can be reproduced is dependent on the database search algorithm used.

Conclusions

We introduced here a simple model to predict the cluster centres of peptide masses. The input parameters of the model can be easily determined for the sequence databases. We studied how these parameters influence the location of cluster centres, concluding that the cleavage specificity of the enzyme used for peptide digestion and the cleavage probability are the main factors. The change of the cluster centre location due to changes in average protein length or due to variability of amino acid frequencies among the databases is relatively small. However, our analysis also illustrates that, due to combinatorial constraints, the

location of the cluster centres for masses smaller than $1000Da$ can differ from the average location.

Based on the model of the peptide mass cluster centres we derived a measure to determine the deviation of an experimental peptide mass from the nearest cluster centre. We used this distance measure to calibrate the peptide peak-lists and to recognise non-peptide peaks. The calibration method, linear regression on peptide rule, is a robust and accurate method to calibrate single peak lists without resorting to internal calibrants. With this method higher calibration precision was obtained in comparison to other calibration methods, which also employ the property of peptide mass clustering.

The same distance measure was used to recognise non-peptide peaks and to remove them from the peak-lists. Due to their removal, an increase of the identification rate of up to 2.5% for the PBMS scoring schema was observed.