

Analysis of sets and collections of Peptide Mass Fingerprint data



Dissertation zur Erlangung des akademischen Grades
des Doktors der Naturwissenschaften (Dr. rer. nat)
eingereicht im Fachbereich Biologie, Chemie, Pharmazie
der Freien Universität Berlin

vorgelegt von
Witold Eryk Wolski
aus Warschau

31.05.2007

Gutachter

- Erster Gutachter : Prof. Dr. Udo Heinemann
(Department of Crystallography, Max-Delbrück-Center for Molecular Medicine)
- Zweiter Gutachter : Prof. Dr. Knut Reinert
(Institut für Informatik, Fachbereich Mathematik und Informatik, Freie Universität Berlin)

I would like to dedicate this work to Pawel Grec and Martin Mätzig.

Acknowledgements

During the four years of my graduate studies, I have had cause to be grateful for the advice, support and understanding of many people. In particular, my supervisors, Knut Reinert and Hans Lehrach, had been a constant source of good ideas and sound advice.

Much advice and many useful discussions were provided by Maciej Lalowski, Johan Gobom, Ralf Herwig, Thomas Kreitler, Peter Martus, Peter Jungblut, Clemens Gröpl and Andreas Döring.

Last but not least I would like to thank Beate Schröder for her understanding, support and true friendship during my work on this thesis.

Abstract

Recent advances in genomics, which outstanding achievements were exemplified by the complete sequencing of the human genome provided the infrastructure and information enabling the development of several proteomic technologies. Currently no single proteomic analysis strategy can sufficiently address the question of how the proteome is organised in terms of numerical complexity and complexity generated by the protein-protein interactions forming supramolecular complexes within the cell. In order to bring a detailed structural/functional picture of these complexes in whole genomes, cells, organelles or in normal and pathological states several proteomic strategies can be utilised. Combination of technologies will bring a more detailed answer to what are the components of certain cellular pathways (e.g.: targets of kinases/phosphatases, cytoskeletal proteins, signalling molecules), how do they interconnect, how are they modified in the cell and what are the roles of several complex components in normal and disease conditions. These types of studies depend on fast and high throughput methods of protein identification. One of the most common methods of analysis is mass spectrometric technique called peptide mapping. Peptide mapping is the comparison of mass spectrometrically determined peptide masses of a sequence specific digest of a single protein or peptide of interest with peptide masses predicted from genomic databases. In this work several contributions to the computational analysis of mass spectrometric data are presented. During the course of my studies I looked at the distribution of peptide masses in sequence specific protein sequence digests and developed a simple mathematical model dealing with peptide mass cluster centre location. I have introduced and studied the methods of calibration of mass spectrometric peak-list without resorting to internal or external calibration samples. Of importance is also contribution of this work to the calibration of data produced in high throughput experiments. In addition, I studied how filtering of non-peptide peaks influences the identification rates in mass spectrometric instruments. Furthermore, I focused my studies on measures of spectra similarity which can be used to acquire supplementary information, increasing the sensitivity and specificity of database searches.

Original Publications

- I **Wolski WE**, Farrow M, Emde AK, Lehrach H, Lalowski M, Reinert K: **Analytical model of peptide mass cluster centres with applications.** *Proteome Sci* 2006, **4**:18.

- II **Wolski WE**, Lalowski M, Jungblut P, Reinert K: **Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants.** *BMC Bioinformatics* 2005, **6**:203.

- III **Wolski WE**, Lalowski M, Martus P, Herwig R, Giavalisco P, Gobom J, Sickmann A, Lehrach H, Reinert K: **Transformation and other factors of the peptide mass spectrometry pairwise peak-list comparison process.** *BMC Bioinformatics* 2005, **6**:285.

Contents

1	Introduction	1
2	Biological Mass spectrometry	6
2.1	Mass spectrometry	6
2.2	Protein mass spectrometry	7
2.3	Peptide Mass Fingerprinting by Matrix-Assisted Laser Desorption Time of Flight Mass Spectrometry	9
2.4	Protein identification by Electrospray ionisation tandem mass spectrometry	14
2.5	Protein identification using Mass spectrometric data and sequence information	20
2.5.1	Feature extraction	21
2.5.2	Database scoring schemes	22
2.5.2.1	Peptide Mass Fingerprinting	22
2.5.2.2	MS/MS peptide ion fragmentation pattern search	24
2.5.3	Validation of search results	28
2.6	Summary	29
3	A mathematical model of the peptide mass rule with applications	30
3.1	Introduction	30
3.2	Methods	33
3.2.1	Data sets	33
3.2.2	Calibration	34
3.2.3	Sequence databases	34

3.2.4	In Silico Protein Digestion	34
3.2.5	Regression analysis	35
3.2.6	Wool and Smilanskys algorithm	36
3.3	Results and Discussion	37
3.3.1	A simple way to predict the peptide mass cluster centres of a protein database	37
3.3.2	Error of the model	45
3.3.3	The type of distribution around the cluster centres	46
3.3.4	Sensitivity analysis	48
3.3.5	A measure of distance to cluster centres	49
3.3.6	Applications	50
3.3.6.1	Linear Regression on Peptide Mass Rule LR/PR	50
3.3.6.2	Filtering of non-peptide peaks using the peptide mass rule	52
4	Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants	57
4.1	Introduction	57
4.1.1	Two novel calibration methods	58
4.1.2	Evaluating the methods	59
4.2	Methods	60
4.2.1	Data sets	60
4.2.2	Describing the Mass Measurement Error and predicting the correct mass	61
4.2.3	Affine mass measurement error model	62
4.2.4	Determining ubiquitous masses and their filtering	62
4.2.5	Standard internal calibration - Alignment to a pre-compiled list of calibration masses	63
4.2.6	Filtering of ubiquitous masses prior to database search	64
4.2.7	Thin-plate spline	64
4.2.8	Linear Regression and Peptide mass Rule algorithm	65
4.2.9	External Calibration	66
4.2.10	Similarity/Quality measures for internal calibration	66

4.2.11	Alignment of a set of peak-list using a Minimum Spanning Tree	67
4.3	Results and Discussion	70
4.3.1	Internal calibration using a pre-calibrated list of calibration masses	70
4.3.2	Finding the optimal multiple peak-list alignment using a modified Minimum Spanning Tree algorithm	72
4.3.3	Determining the calibration model of the sample support using Thin-Plate Spline interpolation	74
4.3.4	The mass measurement error	78
4.3.5	The optimal size of the search window	79
4.3.6	Combining different calibration methods and their comparison	80
4.3.7	The BioConductor package <code>mscalib</code>	84
4.4	Conclusions	84
5	Transformation and other factors of the biological mass spectrometry pairwise peak-list comparison process	87
5.1	Introduction	87
5.1.1	The subtractive analysis technique	88
5.1.2	The pairwise peak-list comparison process	89
5.1.3	Evaluation Framework	90
5.2	Methods	93
5.2.1	Datasets and pre-processing	93
5.2.1.1	PMF-data	93
5.2.1.2	MS/MS data	94
5.2.2	Finding the matching peaks	95
5.2.3	Weighting the missing mass measurement accuracy	96
5.2.4	Non matching peak pairs	97
5.2.5	Binary measures	97
5.2.6	Peak intensity Scaling	101
5.2.7	Measures based on peak intensities and intensity ranks.	102
5.2.8	Computation	104

CONTENTS

5.3	Results and Discussions	105
5.3.1	The number of matches	105
5.3.2	The number of non-matching peaks	107
5.3.3	Peak intensities	108
5.3.4	Peak intensity transformation	108
5.3.5	The factors of the pairwise peak-list comparison	110
5.3.6	The evaluation scores	111
5.3.7	ANOVA of the pairwise peak-list comparison approaches	113
5.3.8	The ANOVA results	113
5.3.9	Dissimilarity measures with small variance and high PAUC scores	116
5.3.10	Intensity transformation and ANOVA	119
5.3.11	The peak-list length	120
5.3.12	Differences between binary and intensity based dissimilarities	121
5.3.13	Weighting of mass measurement accuracy, computing the non-crossing matching and weighting of non-matching peaks	122
5.4	Conclusions	123
6	Conclusions	125
	References	148

List of Figures

2.1	Basic components of a spectrometric instrument: ion source, mass analyser, detector and processing unit.	7
2.2	Schema of Protein Mass Fingerprinting	11
2.3	Soft Laser Desorption Process (1)	13
2.4	Schema of a Time of Flight Spectrometer.	14
2.5	Problems of Protein identification by MALDI-TOF. Top – protein contamination by <i>e.g.</i> keratins can be introduced during the protein separation. Top left – Autoproteolysis products of the protease, contaminating the sample. Bottom left – During the desorption process, matrix molecules are volatilised, ionised and detected allowing the contamination of the spectra. Because of the limited mass measurement accuracy of the MS instrument, a mass measurement error is introduced (bottom).	15
2.6	Schema of protein identification using tandem mass spectrometry. The protein sample (top left) contains proteins (A, B, C). The protein mixture is digested using sequence specific proteases. The <i>peptide mixture</i> is separated using liquid chromatography (LC) technique and MS/MS spectra of peptides are acquired. The peptides are identified by database searches (bottom right). By applying peptide grouping and data analysis (top right) the proteins A, B, C are identified (2).	16
2.7	The electrospray process (1).	17

LIST OF FIGURES

2.8	Principle of tandem mass spectrometry. Top panel: Parent ion spectrum presenting peptide mass peaks of tryptic digest of the protein hormone insulin. Bottom panel: Fragment ion spectrum of the <i>parent peptide</i> mass enclosed in the red square in the top panel with peptide sequence GFFYPTK.	18
2.9	CID induced fragmentation pattern and Fragment ion nomenclature. N-terminal a_2 , b_2 , c_2 ions and C-terminal x_3, y_3, z_3 ions for a five amino acid peptide.	19
2.10	In plot A, B, C, and D the x-axis represents the mass interval between $2230Da$ and $2250Da$, whereas the y-axis shows the intensity. A: Part of a MALDI mass spectrum. Plots B, C, and D show the continuous wavelet transform of the spectrum using a Marr wavelet with different dilation values a (B: $a = 3$, C: $a = 0.3$, D: $a = 0.06$) (3).	21
2.11	Principle of peptide identification using MS/MS data. The experimental spectrum (left) is compared to theoretical spectra (right). A scoring algorithm compares the this two spectra and computes the significance of the similarity.	25
2.12	Partial interpretation of a spectrum. Simplified representation of an MS/MS spectrum for the peptide GFFYTPK. The b-ion ladder is shown in green, while the y-ion ladder is denoted by blue colour. Distances between peaks on the horizontal mass-to-charge (m/z) axis can be used to infer partial sequences of the peptide. This example shows how the partial sequence FFE can be inferred from the y-ion ladder.	27
3.1	The peptide mass rule. Panel A: Scatterplot of $m^{(M)} - m^{(N)}$ against the $m^{(N)}$ mass ($m^{(M)}$ - monoisotopic mass, $m^{(N)}$ - <i>nominalmass</i>). 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.	38

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.	39
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 <i>DHB</i> (2,5-dihydroxybenzoic acid) matrix clusters from the peptide mass cluster centres; dashed line – distance of <i>alphacyano</i> (alpha-Cyano-4-hydroxycinnamic acid) clusters from the peptide mass cluster centres; distance of <i>sinapicacid</i> (3,5-Dimethoxy-4-hydroxycinnamic acid) clusters from peptide mass cluster centres. . .	44
3.4	The monoisotopic mass as an function of the nominal mass. Left panel : $m^{(M)} - m^{(N)} = (\lambda_{RC,pc}^{(m),*} - 1) \cdot m^{(N)}$. Right panel : Difference between $(\lambda_{RC,pc}^{(m),*} - 1) \cdot m^{(N)}$ and $0.00048 \cdot m^{(N)} + 0.029$	45
3.5	Difference between cluster centre computed for the <i>in silico</i> 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.	46
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 $m \in (500 - 530)$, $m \in (1000 - 1110)$, $m \in (2000 - 2200)$ and $m \in (3400 - 3700)$	47
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.	49

3.8	Principle and results of linear regression on peptide rule <i>LR/PR</i> 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.	51
3.9	Schema of non-peptide mass filtering. Abscissae - peptide mass, ordinate – $m \bmod 1$, 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.	53
3.10	Scatter plot : abscissae - peptide mass m_i , ordinate - $m_i \bmod \lambda$ 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.	54
4.1	Modified Dijkstra-Prim minimum spanning tree algorithm. The algorithm starts with vertex s (peak-list) belonging to the peak-list pair with smallest distance (line 1) (the standard algorithm starts with an arbitrary pair). In addition to computing the minimum spanning tree T , the algorithm computes the calibration constants $C(v, s)$ (line 8) and the connection weight $W(u)$ (line 9).	68

4.2	<p>A: Histogram of masses present in the stick spectra in B. In red, marked masses recognised as ubiquitous. B: Stick spectra of five hypothetical peak-lists. Red vertical lines mark the position of ubiquitous masses determined using the histogram in A. C: Single linkage-clustering dendrogram of the peak-lists in B. As dissimilarity the mass measurement range (1500 Da) minus the range enclosed by matching peaks was used. D: Minimum spanning tree.</p>	71
4.3	<p>A: Colour scheme coded peak-list lengths in dependence of the sample support position. Blue dots – <i>interior</i> vertex, Green dots – end vertex, white arrows – connecting edges of the minimum spanning tree. The red hair-cross indicates the peak-list of origin s. B: Colour scheme coded slope coefficient of the mass- dependent calibration function in relation to sample support position. C_1, C_2: Strip chart of the data set for a mass range of 2210 – 2212Da (top) and 842 – 843Da (bottom), including the tryptic autolysis peaks 842.508 and 2211.100Da. Black hair-crosses – masses of peaks before calibration, red circles – masses after calibration. Vertical blue line – the exact position of trypsin autolysis masses 842.508 and 2211.100Da.</p>	74
4.4	<p>A : Colour scheme coded slope coefficients c_1 of the mass measurement error determined by the peptide rule based calibration method. B: The slope coefficient as predicted from the refined samples determined by <i>Thin plate spline</i> with $\lambda = 0.001$. 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 the peptide rule based calibration method, red circles – masses predicted by the thin plate spline calibration method. Vertical blue line – exact position of trypsin autolysis masses 842.508 and 2211.100Da. Dashed red vertical line – mass of the extreme peptide masses after thin plate spline calibration.</p>	76

LIST OF FIGURES

4.5	A: Histogram of pairwise peak-list 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 and thin plate spline calibrated data. B: Strip chart of peak-lists. Grey triangles – masses after thin plate spline calibration, green circles – data after thin plate spline and minimum spanning tree calibration, red circles – data calibrated into the theoretical co-ordinate system, defined by theoretical tryptic autolysis masses (blue vertical lines.)	77
4.6	Stick spectrum of the merged data set of 380 peak-lists. The black vertical lines represent peaks calibrated using the thin plate spline and minimum spanning tree method. Their height equals their intensity. Green line – average mass of all peaks in the region 842 – 843Da (A) and 2210.5 – 2211.6Da (B). The orange vertical lines represent the average mass \pm , the standard deviation of the peak masses in each region. Magenta line – density of peak-masses.	79
4.7	The optimal search window. Comparison of the relative identification rates of internally calibrated data (Y-axis) given a search window size of 0.5Da, 0.2Da, 0.1Da, 0.05Da and 0.02Da, respectively (X-axis). Red – Two Reflex (Pirellula) dataset, Black – Two Ultraflex (<i>Mus Musculus</i>) datasets	80
4.8	Relative identification rate in % (continuous line – left y-axis) and sequence coverage in % (dashed lines - right y-axis). LR/PR – linear regression on peptide rule, IC – two step internal calibration, MST – minimum spanning tree calibration, P – thin plate spline calibration, TPS-IC – thin plate spline calibration and subsequent internal calibration, TPS-MST - thin plate spline calibration and subsequent minimum spanning tree calibration. Panel A - Reflex datasat, Panel B - Ultraflex dataset, Panel C - Average of Dataset (see text for details).	83
5.1	Example of a peak-list stick spectrum for fragment ion MS/MS (top panel) and PMF(bottom panel). X-axis – mass of the peaks, Y-axis – area under the peak.	91

LIST OF FIGURES

5.2	Stick spectrum of two peak-lists X (black lines) and Y (black dot dashed lines). Upper left corner – accuracy of the mass measurement a . <i>A</i> – ambiguous match of five peaks. <i>B</i> – unambiguous match of two peaks. <i>C</i> – peaks not matching.	96
5.3	A – Histogram of the number (bandwidth = 1) of matching peaks for peak-lists chosen from the same cluster (magenta) and from different clusters (green). B – Histogram of the number (bandwidth = 3) of non-matching peaks, if peak-lists were chosen from the same (magenta) or from different clusters (green).	107
5.4	Peak Intensities. A – Histogram of intensities: X-axes – Intensity of log transformed root-means-square scaled peak intensities. Y-axis – Frequency. In grey: Histogram of the peak intensities that do not match a peak in any other peak-lists (peak-lists) <i>within</i> the same cluster (this mass is observed only once in the cluster). In magenta: Histogram of intensities of peaks that do <i>match</i> a peak within any peak-list <i>within</i> cluster (this mass is observed at least twice in the cluster). B – Altman Bland plot of intensities of the matching peaks for peak-lists pairs from <i>within</i> a cluster. C – Altman Bland plot of intensities of matching peaks for peak-lists pairs of <i>between</i> clusters.	109
5.5	Receiver Operator Characteristic curve - The sensitivity (TP-rate) is plotted against $FP = 1 - specificity$ using the number of matching peaks as the discriminatory variable. Red dashed area: sensitivity-PAUC – partial area under the ROC curve for FP-rate $\in [0, 0.1]$. Green dashed area: specificity-PAUC – partial area under the ROC curve for sensitivities $\in [0.9, 1]$	112

5.6 **A:** Boxplot of the sensitivity-PAUC (sensitivity given a FP-rate $\in [0, 0.1]$) itemised according the factors *dissimilarity measure* and θ (weighting of non-matching peaks) for the binary measure based peak-list comparisons. **B:** Boxplot of the factors *scale* (cf. Methods - Scaling) and *measure* of the sensitivity-PAUC (sensitivity given a FP-rate $\in [0, 0.1]$) for intensity measure based peak-list comparisons. The **top** panels show a clip (ZOOM) of the bottom boxplot, indicated by the green horizontal line. X-axis labels: fm – Fowlkes-Mallows statistics, gower – Gower coefficients, hg – Huberts Γ , rmi – relative mutual information, canberra – Canberra distance, simindex – similarity index, manhattan – Manhattan distance, euclidean – Euclidean distance, dotprod – dot-product measure, cov – covariance, soai – sum of agreeing intensities. Scaling: T – total ion count, N – vector length, S – root mean square, R – ranks 117

5.7 Boxplot **A:** Comparison of the sensitivity-PAUCs (computed for FP-rate $\in [0, 0.1]$) computed for the assymmetric binary measures with sensitivity-PAUCs of the symmetric binary measures, in case of the PMF dataset. Boxplot **B:** Comparison of the sensitivity-PAUC (computed for FP-rate $\in [0, 0.1]$) computed for the assymmetric binary measures with the sensitivity-PAUCs of the symmetric binary measures, in case of the MS/MS dataset. 118

5.8 **A:** Boxplot of the specificity-PAUC (specificity given a TP-rate $\in [0.9, 1]$) for the dot-product measure (dotprod) and sum of agreeing intensities (soai). **B** Boxplot of the sensitivity-PAUC (sensitivity given a FP-rate $\in [0, 0.1]$). N – raw intensities, S – square root transformed intensities, L - log transformed intensities, R - intensity ranks. 120

List of Tables

3.1	Masses of Atoms	31
3.2	Protein lengths and amino acid frequencies (one letter code) for nine in the nine databases. <i>length</i> – average protein length in database, <i>reference</i> – database reference; f_i – amino acid frequencies	35
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.	41
3.4	Cleavage sites of proteolytic enzymes (4)	48

LIST OF TABLES

3.5	Results for filtering of non-peptide masses. Columns: <i>Arabidopsis t.</i> , <i>Rhodopirelulla b.</i> , <i>Mus musculus</i> – 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.	55
4.1	Mass Measurement Error. Standard deviation (S_N) observed for the tryptic autolysis peaks 842.508 and 2211.1. Raw data; TPS - thin plate spline calibrated data; TPS-MST - The data, which undergone Thin-Plate Spline (TPS)(pre-processing), followed by Maximum Spanning Tree (MST) calibration	78
4.2	Calibration sequences. LR/PR – linear regression on peptide rule, IC – Internal calibration with two iterations. (Bruker Reflex – mass measurement error window of 450 and 250ppm, Bruker Ultraflex – 250 and 125ppm); MST – minimum spanning tree calibration method computed with an search window of $\pm 0.4Da$; TPS-IC - Pre-processing (thin plate spline calibration) and subsequent internal calibration with a mass measurement error window of 250ppm; TPS-MST - thin plate spline pre-processing and an minimum spanning tree with a search window of $\pm 0.25Da$; . . .	81

LIST OF TABLES

5.1	Number of clusters of given cluster size N . The columns 2 and 3 describe the cluster size in the PMF- and the MS/MS datasets. Number of spectra – number of peak-lists submitted for database search, identified spectra - spectra assigned to a database ID with an either significant probability based Mowse score (PMF-data) or to a peptide sequence with $Xcorr > 2$, and an ion coverage $> 20\%$ (MS/MS-data) given a parent peptide charge $z = 2$. Identified proteins/peptides - the number of uniquely identified proteins or peptides. ^A – approximate number of spectra derived from ion fragments of peptides with charge $z = 2$. ^B – The number of spectra with charge $z = 2$ of the parent ion ($\approx 53\%$ of all identified spectra).	92
5.2	Modified contingency table. $M = \max\{N, c + \theta \cdot (M_{01}^{XY} + M_{10}^{XY}) + M_{11}^{XY}\}$ with N defined by the user and $c = 1$ in case of Hubert's Gamma or $c = 0$ otherwise.	97
5.3	Peptide (PMF) peak-list and peptide fragment ions (MS/MS) peak-list properties. MME – mass measurement error. The rows 1 and 4 provide a <i>five-number summary</i> and the <i>mean</i> of the peak-lists lengths (number of peaks in peak-list) in the dataset. Rows 2,3 (PMF) and 5,6 (MS/MS) provide the <i>five-number summary</i> and the <i>mean</i> of the number of matches observed if comparing <i>within</i> and <i>between</i> cluster peak-lists pairs. Min. - minimum, 1st Qu. - first quartile, 3rd Qu. - third quartile, Max. - maximum	106
5.4	The adjusted R^2 of the model $ \Delta I \sim \bar{I} + \bar{I}^2$ for the raw, squared (Tabb et al. (5)) and log transformed peak intensities. PMF – PMF-data; MS/MS – MS/MS data.	110
5.5	Factors considered in the comparison process and their levels. Column 1 – Factors: identification of factors, Column 2 – Levels: short summary of the levels (For more details please refer to the Methods section). Column 3 – Number: number of levels. Int. – comparisons considering the intensities; Bin. – binary measures.	111

- 5.6 Influence of factors specifying the pairwise peak-list comparison on partial areas under the ROC curve for binary PMF and MS/MS data. For each of the 96 pairwise comparison approaches, sensitivity-PAUC (sensitivity given FP-rate $\in [0, 0.1]$) and specificity-PAUC (specificity given sensitivity $\in [0.9, 1]$) (Figure 5.5) were determined. A partitioning of sums of squares was performed analogously to analysis of variance. Column names: Factors – identification of factors ; df – degrees of freedom (DF, number of factor levels - 1); %SSQ – relative sum of squares ($\%SSQ = SSQ / \sum SSQ$); %MSQ – relative mean sum of squares ($\%MSQ = MSQ / \sum MSQ$), where $MSQ = SSQ / DF$. %MSQ measures the importance of a specific factor for the size of specificity-PAUC and sensitivity-PAUC. \times denotes interactions between factors. measure – distance measure, noncross – non crossing matching, length – alignment length, θ – weight of non-matching peaks, residual – unexplained %SSQ or %MSQ, total – column sum of %SSQ, df, %MSQ. 114
- 5.7 Influence of factors specifying the pairwise peak-list comparison on partial areas under the ROC curve for intensity PMF and MS/MS data. For each of the 2688 pairwise peak-list comparison approaches, sensitivity-PAUC (sensitivity given FP-rate $\in [0, 0.1]$) and specificity-PAUC (specificity given sensitivity $\in [0.9, 1]$) (Figure 5.5) were determined. A partitioning of sums of squares was performed analogously to analysis of variance. Column names: Factors – identification of factors ; df – degrees of freedom (DF, number of factor levels - 1); %SSQ – relative sum of squares ($\%SSQ = SSQ / \sum SSQ$); %MSQ – relative mean sum of squares ($\%MSQ = MSQ / \sum MSQ$), where $\%MSQ = MSQ / \sum MSQ$. %MSQ measures the importance of a specific factor for the size of sensitivity-PAUC and specificity-PAUC. \times denotes interactions between factors. measure – distance measure, noncross – non crossing matching, length – alignment length, θ – weight of non-matching peaks, trans – peak intensity transformation, residual – unexplained %SSQ or %MSQ, total – column sum of %SSQ, df, %MSQ. 115

Nomenclature

2D – PAGE Two Dimensional Polyacrylamide Gel Electrophoresis

mod modulo operator

ANOVA analysis of variance

CID collision induced dissociation

DFT Discrete Fourier Transformation

ESI Electrospray Ionisation

EST Expressed Sequence Tag

FN false negative

FP false positive

HPLC high performance liquid chromatography

m/z mass over charge

MALDI matrix assisted laser desorption/ionisation

MS Mass Spectrometry

MST minimum spanning tree

OO Object-Oriented

PAUC partial area of interest under ROC curve

LIST OF TABLES

PBMS Probability based mascot score

PMF Peptide Mass Fingerprinting

ROC receiver operator characteristic

SDS Sodium dodecylsulphate

TN true negative

TOF time of flight

TP true positive

TPS Thin-Plate Spline

LR/PR peptide rule calibration