Chapter 6

Conclusions

Through our work we have contributed to the computational methods of the protein identification by peptide mapping that comprise of peak detection, recalibration of peptide mass lists, removal of non-peptide peaks, and finally grouping and assigning peptide mass lists to protein sequences.

We have examined the properties of protein sequence databases and derived a mathematical model of the peptide mass rule, which describes the distribution of peptide masses of a peptide mixture generated from a sequence database. We studied how the parameters of the model influence the location of cluster centres, concluding that the cleavage specificity of the enzyme used for peptide digestion and the cleavage probability are the most important factors to accurately predict the location of the cluster centres. The location of the cluster centres can be used to calibrate peptide masses or to remove non-peptide peaks. Calibration is possible because on average the deviation from the cluster centres predicted by our model, of all peaks in a peptide mass list, should be zero. Hence, if systematic deviations from the cluster centres are observed it indicates a mass measurement error. We have called this calibration method "'linear regression on peptide rule"'. It is a robust and accurate method allowing calibrating single peak-lists without resorting to internal calibrants. Using the method we obtained desired calibration precision and reliability, which make this method to be practically applicable.

After calibrating the peak-lists, which employs minimising the distances of all peptide masses to the cluster centres by a linear transformation, the distance to cluster centres can be utilised to detect non-peptide peaks from peptide peak-

lists. Non-peptide peaks are those peaks, which after peak-list calibration strongly deviate from the cluster centres. Due to their removal, the sensitivity as well as the specificity of protein identification by database searches can be achieved. Applying the non-peptide peak filtering increased the identification rate up to 2.5% in case of the Probability based Mascot scoring scheme.

An important part of this work was the development of calibration methods for sets of peak-lists acquired in high throughput experiments. Samples processed in high-throughput experiments exhibit similarities with many other samples because they, for example share the mass spectrometric sample support or the microtitre plate. Samples sharing the microtitre plate are exposed to identical laboratory conditions, what increases the relative reproducibility of experiments within a set. Hence, learning about one sample supplies us with additional information about another samples in the same set. For example, peaks observed in one sample are more likely to be observed in other samples of the same set than in otherwise unrelated samples. Furthermore, the mass measurement error for samples, deposited at the same mass spectrometric instrument and processed at approximately the same time is highly correlated. Based on these observations we have developed two methods for calibration of mass-spectrometric peak-list sets. One method is based on expected similarities due to contaminants, mass spectrometric matrix and peptide peaks of auto- proteolysis products of the protease. The other method explores the correlation of the mass measurement error for closely related peak-lists.

While the methods described in this study significantly improve the calibration of raw data, they do not perform better than other published calibration routines, which reduce the MME to $10ppm$ or below. The novelty of the methods introduced for calibration of set of peak-lists (41) is, that by exploiting similarities between peak-lists we were able to re-calibrate peak-lists what would be not be possible using conventional calibration methods. Hence, by employing our methods a larger fraction of peak-lists in the dataset can be calibrated and more proteins could be identified.

The other method introduced in this work, explores similarities between the peaks content of peak-lists, resulting from similar chemical processing. In a high-throughput setting a restriction enzyme of identical chemical properties is used.

Hence, peak-lists processed in parallel share a unique set of peaks, characteristic for the batch. These peaks can be used to align the peak-lists. To determine the order in which to align the peak-lists we have used the minimum spanning tree (MST) algorithm.

A common property of these two calibration methods is that they performance increases with higher number of samples in the set. We can conclude, that if peak-lists are deposited closer on the sample support, we are able to measure the mass measurement error more precisely increasing the efficiency of the TPS calibration method. Therefore, high density microtitre plates and sample supports are not only rational with respect to the idea of high throughput experiments – maximal utilisation of energy and resources but also help to obtain better calibration results employing the TPS method. Dense excision of spots from 2D-gels not only helps to identify more proteins but also increases the performance of the MST method.

The next step of protein identification was the classification of peptide mass lists, which were calibrated to high precision. This can be achieved by searching protein sequence databases. In order to complement identification by database search, pairwise peak-list comparison can be employed (Chapter 5). In our work we had concentrated on assessing the performance and identifying the factors on which their performance depends, of a large group of pairwise similarity measures. Furthermore, we examined the performance of various measures, which to our knowledge were not used for the pairwise comparison of peak-lists. We have extended the measures to accommodate means to handle properties specific to mass spectrometry, *e*.g. mass measurement error.

The aim of this part of the work presented here was to determine the pairwise peak-list comparison approach with highest sensitivity and specificity for the grouping of spectra. The primary however was to determine which factors studied had the highest effect on the outcome of the clustering, in order to foster the understanding of the pairwise peak-list comparison process. While the first goal could be easily achieved by ranking the various peak-list comparison approaches, the second goal was approached by analysis of variance (ANOVA) techniques. The partial area under the *Receiver Operator Characteristic* (ROC) curve, determined for high sensitivity and specificity values was used as the dependent variable, while the various choices for the comparison process were the factors in the ANOVA.

To examine whether the obtained results can be generalised to various mass spectrometric datasets we based our study on two distinct datasets (PMF and MS/MS data). The results generated for both datasets were similar, providing evidence that the obtained results might be of general interest.

Two factors, namely measure and intensity scaling and their interactions had the highest impact on the intensity based pairwise peak-list comparisons. The combination of the Euclidean distance with vector norm scaling, the Manhattan distance with total ion count scaling and the sum of agreeing intensities with vector length scaling were the best performing measures. The measures suggested by us were so far not used to assess similarities between mass spectrometric peak-lists. A further factor, which can be used to increase the classification performance of the peak-list comparison is the intensity transformation with the log function as a best choice. In case of the MS/MS data we recommend to apply the weighting of mass measurement accuracy and combine it with a decrease of the weight of non-matching peaks ($\theta = 0.5$), as well as to implement the computation of non-crossing matching.

The most important factors for the comparison of the peak-lists using binary measures are the measure, weight of non-matching peaks (θ) and peak-list length N. Symmetric measures with large peak-list length N and a small weight of non-matching peaks ($\theta = 0.5$) performed best for MS/MS data, while asymmetric measures were the most useful during a comparison of PMF data. A further possible direction to enhance measures of pairwise peak-list dissimilarity would be to combine them with methods that model peak-list properties *i.e.* peptide fragmentation patterns (45).

The recommended pairwise peak-list comparison approaches can be used as predictive functions of *within* and *between* cluster associations of mass spectrometric peak-list pairs. However, the best value of the discriminatory variable still needs to be determined. This can be achieved, for example, by the use of ROC curves combined with cross validation analysis, but will require a dataset where the identities of the peak-lists are known *a-priori*.