DOI 10.1002/pmic.200700057 *Proteomics* 2007, *7*, 3470–3480

RESEARCH ARTICLE

SuperHirn – a novel tool for high resolution LC-MS-based peptide/protein profiling

Lukas N. Mueller[1, 2], Oliver Rinner[1], Alexander Schmidt[1], Simon Letarte[3], Bernd Bodenmiller[1], Mi-Youn Brusniak[3], Olga Vitek[4], Ruedi Aebersold [1, 2, 3, 5] and Markus Müller[1]

[1] Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
[2] Competence Center for Systems Physiology and Metabolic Disease, ETH Zurich, Zurich, Switzerland
[3] Institute for Systems Biology, Seattle, WA, USA
[4] Department of Statistics, Purdue University, West Lafayette, IN, USA
[5] Faculty of Science, University of Zurich, Zurich, Switzerland

Label-free quantification of high mass resolution LC-MS data has emerged as a promising technology for proteome analysis. Computational methods are required for the accurate extraction of peptide signals from LC-MS data and the tracking of these features across the measurements of different samples. We present here an open source software tool, *SuperHirn*, that comprises a set of modules to process LC-MS data acquired on a high resolution mass spectrometer. The program includes newly developed functionalities to analyze LC-MS data such as feature extraction and quantification, LC-MS similarity analysis, LC-MS alignment of multiple datasets, and intensity normalization. These program routines extract profiles of measured features and comprise tools for clustering and classification analysis of the profiles. *SuperHirn* was applied in an MS1-based profiling approach to a benchmark LC-MS dataset of complex protein mixtures with defined concentration changes. We show that the program automatically detects profiling trends in an unsupervised manner and is able to associate proteins to their correct theoretical dilution profile.

Received: January 22, 2007
Revised: June 26, 2007
Accepted: June 26, 2007

Keywords:
Label-free quantification / LC-MS / Protein profiling / Quantitative proteomics

1 Introduction

The accurate identification and quantification of proteins in complex biological mixtures is fundamental for molecular systems biology approaches. LC coupled to MS (LC-MS) in combination with shotgun tandem mass spectrometry (MS2) yields the molecular composition of complex peptide mixtures generated from protein digests (reviewed in ref. [1]). However, only a small subset of peptide features present in the sample are selected for fragmentation analysis leading to a systematic undersampling of peptide identifications in conventional shotgun LC-MS experiments [2]. This effect is caused by various factors such as ion selection bias of the mass spectrometer towards high abundance peptides, co-eluting MS2 precursor ions, and differences in peptide fragmentation efficiency resulting in low quality MS2 spectra [3]. Therefore, quantitative analysis of peptide abundance values based exclusively on peptides identified by their fragmentation ion spectra is poorly reproducible and often spans a narrow dynamic range.

An alternative to an MS2-based approach is the analysis of peptide ion currents directly at the MS1 level that allows mapping of extracted peptide features across LC-MS experiments. For this purpose, feature extraction routines detect peptide signals in noisy background and compute the ion counts of selected peptides (SIC) without relaying on MS2 peptide identification information. Even though the comparison of absolute SIC values is complicated by experi-

Correspondence: Lukas N. Mueller, Institute for Molecular Systems Biology – ETHZ,ETH Hönggerberg, HPT C 75 Wolfgang Pauli-Str. 16, Zurich 8093, Switzerland
E-mail: Lukas.Mueller@imsb.biol.ethz.ch
Fax: +41-44-633-10-51

Abbreviations: AMRT, accurate mass retention time pairs; **APML,** annotated putative peptide markup language; **FT-LTQ,** Fourier transformed-LTQ mass spectrometer; **MS2,** tandem mass spectrometry

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

mental variation, the application of peptide signal intensity normalization methods allows restoring the linear correlation between the original peptide concentration and the measured signal intensity [4]. These properties offer an alternative to costly isotopic labeling strategies where chemically modified peptides are compared within the same LC-MS experiment [5–8]. While quantification based on stable isotope labeling seems to be more accurate [9], it involves more extensive sample processing if more than two (or four in the case of iTRAQ [6]) samples are compared [10].

Various retention time (T_R) normalization methods [11, 12] have been developed to correct chromatographic variability between LC-MS experiments and permit the use of the retention time dimension as an additional constraint in the assessment of peptide similarity. High mass precision mass spectrometers in combination with effective intensity and retention time normalization methods therefore offer the possibility to unambiguously map peptides across different LC-MS patterns without the use of MS2 information [13]. Existing software suites combine these functionalities and have been used in comparative studies to identify differentially expressed features between two or more sample states [2, 13–15]. On the other hand, protein profiling has shown to be a powerful method for the classification of proteins according to their concentration profiles across biological samples. While Andersen *et al*. [16] applied clustering of MS2 identified peptide profiles from centrifugation fractions in combination with localization information to identify components of the centrosome, there is currently no freely available software that implements a combined MS1 and MS2-based profiling and profile processing method (see Table S1 in Supporting Information for an overview of existing open source software tools).

We present a novel software package, *SuperHirn*, for the label-free quantitative analysis of data generated by a Fourier-transform ESI mass spectrometer. *SuperHirn* is a platform for LC-MS data processing comprising various newly developed and adopted functionalities to detect and track features in LC-MS patterns, combine these by a multiple LC-MS alignment process into a *MasterMap*, normalize feature intensities across samples, and analyze feature profile trends by clustering analysis. This profiling approach was applied to extract the main profiling trends from a benchmark dataset of six standard proteins in a complex sample background. Subsequently, the obtained profile clusters are used to identify peptides and proteins with a statistically significant correlation to the theoretical protein concentrations.

2 Materials and methods

2.1 Sample preparations

The tryptic digests of the six standard nonhuman proteins horse myoglobin (Swiss-Prot identifier MYG HORSE), bovine carbonic anhydrase (Swiss-Prot identifier CAH2 BOVIN), horse Cytochrome *c* (Swiss-Prot identifier CYC HORSE), chicken lysozyme (Swiss-Prot identifier LYSC CHICK), yeast alcohol dehydrogenase (Swiss-Prot identifier ADH1 YEAST), and rabbit aldolase A (Swiss-Prot identifier ALDOA RABIT) were purchased from Michrom Bioresources and 500 pmols of each protein sample was resuspended in 40 μL of 0.4% formic acid. The background sample was prepared by protein digestion of 80 μL aliquots of bulk serum (Sigma–Aldrich Chemie, Buchs, Switzerland) and subsequent selective enrichment of *N*-glycosylated peptides [17, 18]. Aliquots were vacuum-dried, resuspended in 20 μL of 5% ACN, 0.1% formic acid, and pooled. Subsequently, one volume (2 μL) of standard protein dilution was mixed with four volumes (8 μL) of the *N*-glycosylated peptides and ten volumes (20 μL) of 5% ACN, 0.1% formic acid solution.

2.2 LC-MS/MS analysis

LC-MS analysis of the dilution samples was performed on a Fourier transformed-LTQ mass spectrometer (FT-LTQ) (Thermo Electron, San Jose, CA), which was connected to an electrospray ionizer. The Agilent chromatographic separation system 1100 (Agilent Technologies, Waldbronn, Germany) was utilized for peptide separation, where the LC system was connected to a 10.5 cm fused-silica emitter of 150 μm inner diameter (BGB Analytik, Böckten, Switzerland). The columns were packed in-house with Magic C18 AQ 5 μm resin (Michrom BioResources, Auburn, CA, USA). The Agilent autosampler was utilized to load samples at a temperature of 6°C. A linear gradient from 95% solvent A (0.15% formic acid) to 30% solvent B (2% water in ACN, 0.15% formic acid) was utilized for peptide separation over 30 min at a constant flow rate of 1.2μL/min. After every sample injection, 200 fmol of the standard peptide (Glu1)-fibrinopeptide B (#F3261, Sigma–Aldrich Chemie, Buchs, Switzerland) were injected in the LC-system to prevent crosscontamination and to monitor the performance of the LC system and the mass spectrometer.

The data acquisition mode was set to obtain MS1 scans at a resolution of 100.000 FWHM (at 400 m/z) followed by MS2 scans in the linear IT of the three most intense MS1-peaks (total cycle time was approximately 1 s). Only signals exceeding 150 counts were chosen for MS2-analysis and then dynamically excluded from triggering MS2-scans for 15 s. The mass spectrometer was set to accumulate 10^6 ions for MS1-scans over no more than 500 ms and 10^4 ions over a maximum of 250 ms for MS2-scans. To increase the efficiency of MS2 attempts, the charge state screening modus of the mass spectrometer was enabled to exclude unassigned or singly charges ions.

2.3 MS2 peptide assignments

Obtained MS2 scans were searched against a human IPI protein database (v.3.15) containing the protein sequences of

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

the six standard proteins. The SORCERER-SEQUEST (TM) v3.0.3 search algorithm run on the SageN Sorcerer (Thermo Electron) was used as a search engine. *In silico* trypsin digestion was performed after lysine and arginine (unless followed by proline) tolerating two missed cleavages in fully tryptic peptides. Database search parameters were set for carboxyamidomethylation (+57.021464 Da) of cysteine residues as a fixed modification. Search results were evaluated with the trans proteomic pipeline (TPP) [19] using the Peptide Prophet (v3.0) [20].

2.4 Program implementation

The software *SuperHirn* is programmed in *C++* and the source code together with detailed documentation material is freely available on http://tools.proteomecenter.org/SuperHirn.php. *SuperHirn* was tested on Linux and Mac OS X platforms and its workflow is illustrated in Fig. 1. An MS1 feature extraction routine is performed on the input LC-MS raw data and MS2 peptide identifications from a database search are associated with detected MS1 features (Fig. 1a). An LC-MS similarity score is then computed for every pair of LC-MS runs by LC-MS alignment (Fig. 1b) and LC-MS similarity analysis (Fig. 1c). The similarity analysis is used to construct an alignment topology (Fig. 1d), which determines the order in which LC-MS runs are combined in the multiple LC-MS alignment (Fig. 1e). The constructed *MasterMap* represents a framework for downstream data analysis (Fig. 1f). Throughout the whole analysis process as illustrated in Fig. 1, intermediate results such as the list of detected MS1 features, LC-MS similarity scores, the constructed *MasterMap*, *etc.* are stored in XML formatted text files. Specifically, *SuperHirn* is compatible to preprocessed LC-MS data from other software tools *via* the newly developed annotated putative peptide markup language (APML) XML format. APML allows to combine functionalities of different LC-MS analysis suites and to exchange processed data between different tools. Please visit http://tools.proteomecenter.org/Corra/corra.php for more information regarding the schema and documentation of APML. All program parameters are described in the Results in Supporting Information.

2.5 Data preprocessing

Data preprocessing consists of the extraction of MS1 features from mzXML [21] formatted LC-MS raw data and the annotation of detected MS1 features with MS2 peptide identifications in pepXML format [19]. While all program routines (Figs. 1b–f) are developed in a generic mode to work with data acquired from different types of mass spectrometers, we present here a feature extraction routine developed for high precision mass spectroscopy data as obtained from an FT-LTQ mass spectrmeter.

The feature detection algorithm extracts peptide signals in high mass-accuracy MS1 scans by a 2-D filter in the mass-to-charge (m/z) and retention time dimension. In the m/z dimension, it centroids raw peak data and reduces m/z signals of an MS1 scan to the corresponding monoisotopic masses, along with the charge state (z) and an integrated intensity value of the detected monoisotopic peak. Finding the monoisotopic peaks and charge states can be formulated as a pattern matching task [22], where the measured signals are compared to isotopic templates. Here we used a refined method, which allows a fast calculation of the monoisotopic masses and which also works for overlapping isotopic patterns. The intensity of a monoisotopic peak is then defined as the total intensity of the fitted isotopic pattern (see the Methods and Figs. S1 and S2 in Supporting Information).

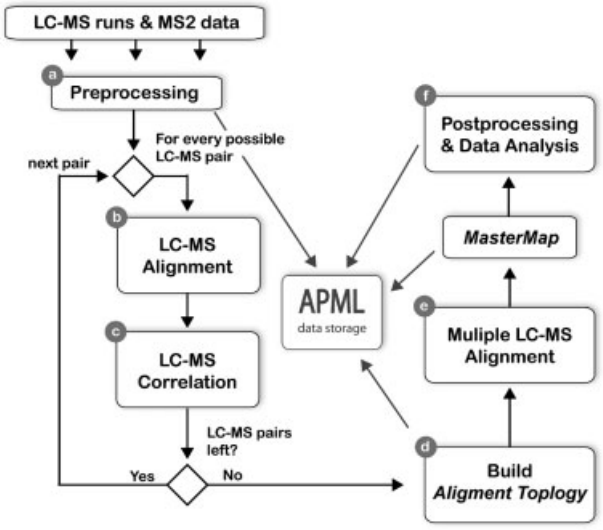

Figure 1. Schematical overview of the program workflow. LC-MS data are preprocessed to extract MS1 features, which are then combined with available MS2 peptide identifications (a). In order to create an alignment topology, LC-MS alignment (b) and LC-MS similarity analysis (c) are performed for every LC-MS pair. The obtained alignment topology (d) guides the multiple LC-MS alignment to construct the *MasterMap* (e). Downstream data analysis is then preformed on this *MasterMap* (f). Preprocessed data are stored in XML files based on the APML format.

The second filter is applied along the T_R dimension and clusters monoisotopic masses of the same charge state and m/z value using an m/z tolerance ($\Delta m/z$) of 0.005 Da. Such monoisotopic clusters represent MS1 features defined by m/z, T_R, and z [23] and are used to compute feature parameters such as retention time of peak apex/start/end, total feature area (V), *etc.* MS2 peptide identifications in pepXML format [19] are then related to their corresponding MS1 feature by retention time, charge state, and the theoretical mass of the peptide sequence. The theoretical mass was preferred to the measured precursor mass to avoid calculation errors of the mass spectrometer operating software in the calculation of the monoisotopic precursor mass. Initially, an MS2 peptide identification is assigned to an MS1 feature if its m/z and T_R difference falls within the m/z and T_R tolerance window. In the cases where more than one peptide identification

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

passes this filtering, the peptide identification with smallest m/z and T_R difference is selected. In general, over 95% of all high probability MS2 peptide identifications (peptide prophet probability >0.9 [20]) were matched to an MS1 feature, while 10% of all detected MS1 features were identified by MS2.

2.6 Retention time normalization by LC-MS alignment

Fluctuations in the LC separation between different LC-MS experiments often lead to a low T_R reproducibility and complicate the usage of the retention time dimension for the tracking of peptide features. *SuperHirn* utilizes a modified $C++$ implementation of the accurate mass retention time pairs (AMRT) method [24] to normalize T_R across LC-MS runs. Initially, all common MS1 features (F_{Common}) of two LC-MS runs A and B are identified within a wide m/z and T_R tolerance window $W_{m/z,\,T_R}$ ($\Delta m/z = 0.05$ Da and $\Delta T_R = 5$ min). The retention time differences $T_{DiffR,\,Obs}$ of F_{Common} (Fig. 2, dark gray dots) follow a trend reflecting the retention time fluctuations between LC-MS runs A and B. The robust smoothing method LOWESS [25] was applied to fit a model $\Delta T_{DiffR,\,Pred}(T_R)$ into the observed retention time differences $T_{DiffR,\,Obs}$. Problematic for the AMRT method are peptides that elute over a wide retention time range and lead to several distinct MS1 feature signals. The counterpart feature in the other LC-MS pattern is therefore matched to several of these features causing multiple $T_{DiffR,\,Obs}$ and a disturbed model of the T_R-shift (Fig. 2, blue dashed line). Filtering out $T_{DiffR,\,Obs}$ values originating from such multiple MS1 features matches (Fig. 2, light gray dots) leads to an improved $\Delta T_{DiffR,\,Pred}(T_R)$ (Fig. 2, green dashed line).

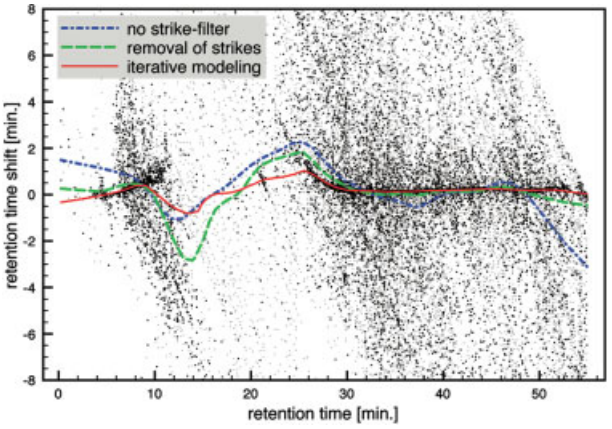

Figure 2. Retention time normalization by LC-MS alignment. Retention time shifts of common features are plotted according to their retention time (dark gray dots) and used to build a model of the retention time shift (blue dashed line). Removing observed shifts from MS1 features matching to multiple common features (light gray dots) leads to a better model (dashed green line). This procedure is performed iteratively to further improve the quality of the model (bold red line).

The implementation of AMRT alignment method was modified to perform the previously described fitting procedure in an iterative manner. Using a sliding window, the local mean of the deviations of $T_{DiffR,\,Obs}$ from $T_{DiffR,\,Pred}$ at a given T_R is computed separately for $T_{DiffR,Obs}$ smaller/bigger than $T_{DiffR,\,Pred}$. These mean deviations $E_{Diff}(T_R)_{UP}$ and $E_{Diff}(T_R)_{DOWN}$ are used as a cutoff to select $T_{DiffR,\,Obs}$ for the next fitting iteration. This procedure is repeated as long as $T_{DiffR,\,Obs}$ are larger than a tolerance value ΔT_R of 0.5 min (Fig. 2, red line). Applying the $T_{DiffR,\,Obs}$ filter and iterative LC-MS alignment leads to an improved correction of retention time values between common MS1 features (see Table S2 in Supporting information).

2.7 Assessment of LC-MS similarity

In order to assess the similarity of two LC-MS runs, we developed a novel strategy to assigns LC-MS similarity scores S_{SIM} to pairs of LC-MS runs. The LC-MS similarity scoring schema is based on the overlap of MS1 features and the reproducibility of their intensities values. Two LC-MS runs are initially subjected to LC-MS alignment and the common features are extracted by comparing their m/z, T_R, and z coordinates ($\Delta m/z = 0.01$ Da and $\Delta T_R = 0.5$ min). A first assessment of LC-MS similarity is the calculation of a feature overlap score $S_{Overlap}$ according to Formula 1 ($N^{A,B}_{Common}$ number of common features, $N_{LC\text{-}MS\ A/B}$ features in LC-MS runs A and B, respectively). For the evaluation of peak area reproducibility, common MS1 features are ranked according to their intensity V separately for each LC-MS run and the robust Spearman correlation coefficient [26] is utilized to compute the intensity score S_{INT} (Formula 2, $R_{A/B,x}$: intensity rank of common feature x in LC-MS runs A and B, respectively). $S_{Overlap}$ and S_{INT} are combined into a final LC-MS similarity score S_{SIM} (Formula 3), which is close to 0 for unrelated runs and approximates 1.0 for highly similar LC-MS runs.

$$S_{Overlap} = \frac{2N^{A,B}_{Common}}{N_{LC-MS\ A} + N_{LC-MS\ B}} \quad (1)$$

$$S_{INT} = \frac{\sum_{x=1}^{F_{Common}} (R_{A,x} - \bar{R}_{A,x})(R_{B,x} - \bar{R}_{B,x})}{\sqrt{\sum_{x=1}^{F_{Common}} (R_{A,x} - \bar{R}_{A,x})^2}\sqrt{\sum_{x=1}^{F_{Common}} (R_{B,x} - \bar{R}_{B,x})^2}} \quad (2)$$

$$S_{SIM} = S_{Overlap} S_{INT} \quad (3)$$

2.8 Multidimensional LC-MS alignment

SuperHirn implements a new muliple LC-MS alignment strategy, which is performed in a fully automated manner without the requirement to define a reference LC-MS run. The computational steps of the alignment process can be grouped into the construction of an alignment topology and

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

the actual multiple alignment process. Analogous to progressive sequence alignment methods (reviewed in ref. [27]), the program *SuperHirn* constructs an hierarchical alignment topology based on the similarity of LC-MS runs defining the order in which the individual LC-MS runs are combined into a *MasterMap*. For every possible pair of LC-MS runs, an LC-MS similarity score S_{SIM} is computed and stored in a matrix. The constructed matrix is subjected to an Unweighted Pair Group Method using Arithmetic Mean [28] hierarchical clustering, where the inverse of the computed S_{SIM} is used as intercluster distance. Hierarchical clustering transforms the matrix into a binary tree structure (alignment topology). Using an LC-MS similarity-based alignment topology, LC-MS runs of poor quality in a dataset are automatically grouped into distance branches of the tree structure and are aligned – if not excluded at all by the user – in a late iteration of the alignment process. Therefore, this strategy represents also a robust method for the alignment of dataset containing LC-MS runs of low quality.

The LC-MS runs are combined during the multi-dimensional LC-MS alignment into a general repository designated *MasterMap*, which stores all aligned MS1 features together with their corresponding MS2 identifications and intensity profiles. *SuperHirn* sequentially searches the two most similar LC-MS runs in the alignment topology and replaces them by a newly merged LC-MS run. This procedure is repeated until only one run remains in the alignment topology. The LC-MS merging process starts with the alignment of an LC-MS pair to remove retention time fluctuations. Subsequently, common MS1 features are searched in the counterpart run within a user defined $\Delta m/z$ and ΔT_R tolerance window ($\Delta m/z = 0.01$ Da and $\Delta T_R = 0.5$ min). In a second step, ambiguities of features, which map to more than one counterpart feature are resolved by selecting the counterpart feature which displays the smallest difference in the m/z and T_R dimension. Common MS1 features are associated to their counterpart as a whole C++ object so that no information of the original feature is lost. MS1 features, which are not present in both runs, are simply represented as a new instance in the merged LC-MS run. A major benefit of the MS1 feature mapping across LC-MS runs is the possibility to exchange MS2 information between common features (MS2 continuation). *SuperHirn* transfers MS2 information from one aligned feature to another if the latter has not been already identified by a high quality peptide assignment (peptide prophet probability > 0.9).

2.9 MS1 feature intensity normalization

Experimental artifacts such as differences in the loading volumes between samples or peptide ionization variations decouple feature intensities from the original peptide concentrations. Therefore, absolute feature intensity values V need to be normalized between acquired LC-MS runs to ensure an accurate label-free quantification [4]. *SuperHirn* uses a modified version of the central tendency normalization method [29], which was originally developed for MicroArray data. The method extracts from the *MasterMap* MS1 features, which were mapped across all LC-MS runs $1...n$, and calculates for every aligned feature the average feature intensity V_{Av}. For every LC-MS run j, the ratio between the intensity of feature i aligned across all n runs and V_{Av} is calculated and averaged over all aligned features. The procedure is then repeated for each run $1...n$ yielding LC-MS specific intensity correction factor. In general, the early and late eluting phase in the chromatography is dominated by noise signals due to the initial equilibration of the LC column and the final washing at the end of the chromatography. Therefore, the normalization procedure is performed locally on retention time segments using a sliding T_R-window to reduce the effect of noisy signals from the early and late eluting LC phase on the normalization of other retention time segments.

Since in some cases (*e.g.*, fractionation experiments) the number of detected MS1 features matched across all LC-MS can be small, this normalization procedure would be biased to only a small subset of features. Therefore, the normalization procedure of *SuperHirn* is performed in an iterative manner starting with MS1 features aligned across the most similar LC-MS runs in the alignment topology and then continuously renormalizes the runs of two branches (i,j) in the tree structure. At every iteration step, an LC segment specific normalization coefficient is computed from features mapped across all N runs of two branches (i,j). To allow some tolerance for features mapped only across some of the runs in one branch, the ratio r of how many times a feature was mapped to the maximal number of possible matches across the runs in a branch is used as a user defined threshold (here $r = 1.0$). The weighted average feature intensity V_{Av} is then computed for an aligned feature according to Formula 4 where N_i and N_j are the number of times a feature was mapped across the LC-MS runs of branches i and j, respectively. The ratio of the average intensity of mapped features in branches i and j to V_{Av} yields then the segment specific normalization coefficient for all features within the LC-MS runs of branches i and j, respectively. This ensures that LC-MS runs with low similarity or small feature overlap to other LC-MS run will not disrupt the normalization process even if the number of aligned MS1 features across all runs is small.

$$V_{Av} = \frac{1}{N_i + N_j} \frac{N_i \sum_{k=1}^{N_i} V_k}{N_j \sum_{k=1}^{N_j} V_k} \quad (4)$$

2.10 Protein profiling analysis

The constructed *MasterMap* with normalized feature intensity values serves as a starting point for the profiling analysis. The profiling comprises a series of computational steps, which are grouped and explained in details in the following

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

three sections: (i) the construction of MS1 feature profiles, (ii) the detection of naturally occurring profiles trends by *K-means* clustering, and (iii) the evaluation of constructed peptide and protein profiles to theoretical abundance profiles. All three steps are specifically performed on the MS1 level and only the construction of protein profiles requires high quality MS2 information. The theoretical profile of interested will be referred in the text as target profile. While in this study the target profiles are obtained from the protein dilution schema, in general enzymatic activity profiles, concentration profiles of markers, *etc.* can be used as target profiles.

2.10.1 Construction of feature profiles

Feature profiles are built from the normalized intensities V of aligned MS1 features. These profiles are further normalized by the sum of all intensities within a profile to obtain values scaled between 0 and 1. A special case is missing data points in a feature profile, which can be caused by several factors (true absence of a peptide in the biological sample, missed detected MS1 feature, incorrect feature alignment, etc.). While it is a common strategy to replace missing profiling data points using interpolation methods, this leads to systematic errors when features are truly missing or their intensities show large variations between neighboring LC-MS runs. In this profiling analysis, missing profile points are set to zero values and no profile smoothing method is applied. To assess the profile similarity of a pair of MS1 features (i,j), a profile score ($S_{Profile}$) based on the Manhattan distance [30] is calculated. $S_{Profile}$ is computed from the sum of differences between the normalized profile values $I_{i,n}$ and $I_{j,n}$ over all profile points divided by the profile length ($N_{LC\text{-}MS}$) (Formula 5). The profile score is confined between 0 and 1 where a high profile correlation is reflected in an $S_{Profile}$ close to zero.

$$S_{\text{Profile}} = \frac{1}{N_{\text{LC-MS}}} \sum_{n=1}^{N_{\text{LC-MS}}} \left| I_{i,n} - I_{j,n} \right| \quad (5)$$

2.10.2 K-means clustering of MS1 feature profiles

The popular *K-means* clustering method [31] is used to group every constructed feature profile. The *K-means* algorithm was chosen due to its transparency, ease of implementation, and its computational complexity $O(N_{Elements} K_{Clusters})$ allowing its application to large datasets (reviewed in ref. [32]). The starting $K_{Clusters}$ cluster centers are randomly chosen from the input features profiles and the clustering cycle is repeated until all cluster centers K reach convergence or a maximal number of iterations is achieved ($N_{Iteration} = 500$). Each built cluster is stored and subsequently used for targeted profiling analysis. A crucial factor in *K-means* clustering is the number of start cluster centers $K_{Clusters}$. We chose the Gap statistic to estimate the optimal number of *K-means* start clusters [33].

2.10.3 Construction of protein profiles

Initially, MS1 features with identical neutral molecular mass and present in different charge states are grouped together to build deconvoluted peptides. Consequently, a deconvoluted peptide represents a set of MS1 features which originate from the same peptide but are present in different charge states. The peptide sequence and the protein belonging of deconvoluted peptides is subsequently inferred from its associated features containing high quality MS2 information (peptide prophet probability >0.9). The protein identifier is then used to catalog deconvoluted peptides into their corresponding proteins. For this analysis, only deconvoluted peptides with a peptide sequence mapping to exactly one protein, are included in subsequent protein profiling analysis. Protein consensus profiles are built by robustly averaging over normalized MS1 feature profiles of all deconvoluted peptides associated to a given protein [16]. After the construction of a protein consensus profile, the protein belonging of a deconvoluted peptide is assessed by its profile correlation to the protein consensus profile. Profile scores between a deconvoluted peptide and protein consensus profile are computed as previously described and the Dixon outlier detection routine ($\alpha = 0.05$) [34] is used to discard outliers. The protein consensus profile is then recalculated and the procedure is repeated until no more outliers are remaining (see Material and Fig. S6 in Supporting Information).

2.10.4 Protein profile correlation to the target profile

To evaluate the correlation between deconvoluted peptides and the target profile, profile scores between all deconvoluted peptides and the target profile are calculated to build a distribution of scores. This distribution is then modeled by a mixture of two Gaussian curves which are used to compute a profile probability ($P^{Pep}_{Profile}$) of a true correlation (+) between the deconvoluted peptide profile and the target profile (Formula 6, $p(+/-)$ are the *a priori* probabilities and $p(S_{Pep.}|+/-)$ the conditional probabilities).

$$P^{\text{Pep}}_{\text{Profile}} = \frac{p(S_{\text{Pep}}|+)p(+)}{p(S_{\text{Pep}}|+)p(+) + p(S_{\text{Pep}}|-)p(-)} \quad (6)$$

The parameters of the two Gaussian distributions are found by means of the expectation maximization (EM) procedure [35]. Protein profile probabilities ($P^{Pep}_{Profile}$) are then calculated from obtained peptide profile probabilities according to Formula 7. $P^{Pep}_{Profile}$ defines the probability that at least one of the deconvoluted peptide profiles of a protein shows a true correlation to the target profile [36]. Subsequently, proteins are ranked by their profile probability reflecting their correlation to the target profile.

$$P^{\text{Prot}}_{\text{Profile}} = 1 - \prod_{i=1}^{N_{\text{b peptides}}} (1 - P^{\text{Pep i}}_{\text{Profile}}) \quad (7)$$

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

3 Results

3.1 A defined profiling benchmark dataset

We used a dilution mixture of six nonhuman purified proteins to evaluate the performance of *SuperHirn*. This dataset consists of two-fold dilution series of the six proteins myoglobin, carbonic anhydrase, cytochrome *c*, lysozyme, alcohol dehydrogenase, and aldolase A spiked into a complex sample background of human peptides isolated by solid-phase *N*-glycocapture from serum. The dilutions were designed and performed according to statistical principles [37] spanning a dynamic range of two orders of magnitude from 25 to 800 fmol injected (Table 1). Each dilution step was analyzed in triplicates on an FT-LTQ instrument and obtained MS2 scans were searched against the human sequence database containing the six nonhuman proteins of the sample.

Table 1. Dilution outline of standard proteins in the benchmark dataset. The dilution schema of the six purified proteins horse myoglobin, bovine carbonic anhydrase, horse cytochrome *c*, chicken lysozyme, yeast alcohol dehydrogenase, and rabbit aldolase A is shown

	Protein injected (fmol) *per* sample					
Protein name	1	2	3	4	5	6
Myoglobin	800	25	50	100	200	400
Carbonic anhydrase	400	800	25	50	100	200
Cytochrome *c*	200	400	800	25	50	100
Lysozyme	100	200	400	800	25	50
Alcohol dehydrogenase	50	100	200	400	800	25
Adolase A	25	50	100	200	400	800

3.2 Construction of the MasterMap

The acquired 18 LC-MS runs were subjected to the MS1 feature extraction routine and in average 15 887 ± 213 MS1 features were detected *per* LC-MS run. A subsequent similarity analysis of the preprocessed LC-MS runs showed a consistently high level of data reproducibility (Fig. S3 in Supporting Information). A *MasterMap* was then created from the six dilution steps for every replicate set and MS1 feature intensities were normalized as previously described. *SuperHirn* computation was performed on an AMD Opteron Processor 250 (4GB RAM, 1GB disk). The feature extraction routine required around 1 min CPU time *per* replicate set (six LC-MS runs), while the construction of the *MasterMap* required 34 min CPU time. All raw mzXML files, MS2 peptide identifications, individual preprocessed LC-MS runs, and constructed *MasterMaps* are available on the *SuperHirn* website.

The mass accuracy ($\Delta m/z$) of the MS instrument is a crucial factor for the mapping of MS1 features during the multiple LC-MS alignment process. We therefore estimated the probability of falsely aligning dissimilar MS1 features as a function of $\Delta m/z$. Retention time values of MS1 features were shifted artificially by 3 min separately for every dilution step and the multiple LC-MS alignment was carried out by using a narrow T_R-window ($\Delta T_R = 0.5$ min). Therefore, no true common MS1 features between two LC-MS runs can be found since the artificially introduced T_R shift exceeds the T_R tolerance and extracted common MS1 features are due to random matches. The procedure was repeated for several $\Delta m/z$ of 0.005, 0.01, 0.05, 0.1, 0.5, 1.0, and 2.0 Da and the obtained distributions of the feature counts were almost binomial (Fig. S4 in Supporting Information). The likelihood p of wrongly matching MS1 features strongly depends on the mass accuracy and is small for the utilized high mass to charge tolerance ($\Delta m/z = 0.01$ Da, $p = 0.029$).

As discussed by Li *et al.* [2], the peptides identified by MS2 are a random draw of a small subset of the total list of peptide present in the sample (MS2 under sampling). High abundance peptides dominate the MS2 analysis whereas identified peptides at low abundance are under represented. To increase the peptide information content, MS2 continuation utilizes MS2 information, which is present in one LC-MS run to annotate the corresponding features in the other runs. All 18 LC-MS runs of the benchmark profiling dataset were combined twice into a *MasterMap* using either the conventional shotgun MS2 assignment or the MS2 continuation method. Figure 3 shows the distribution of identified MS1 features in the *MasterMap* in relation to the number of times they were detected in every individual LC-MS run (Fig. 3, green line). Clearly, only a small fraction of MS1 features are identified independently in every LC-MS run by conventional shotgun sequencing. This observation illustrates the low reproducibility of MS2 sampling in shotgun mode, which makes it difficult to compare different LC-MS runs on the MS2 level only. In contrast, the mapping of peptides on the MS1 level significantly increases the coverage of peptide

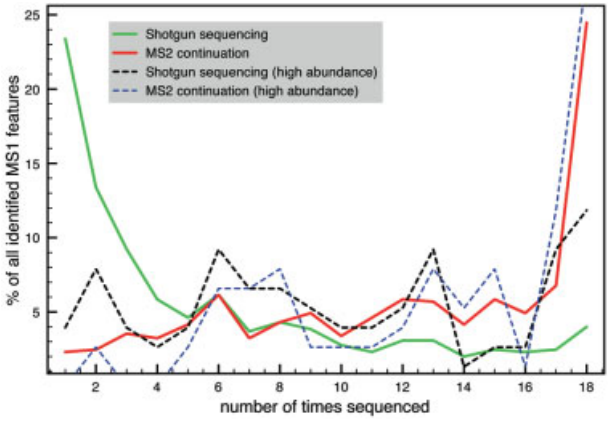

Figure 3. MS2 shotgun sequencing *vs.* MS2 continuation. The distribution of feature counts in all 18 runs is shown for all (green line) and only high abundance features (black dashed line). The same analysis is performed using MS2 continuation for all (red line) or high abundance features (blue dashed line).

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

identifications (Fig. 3, red line). In addition, the same analysis was performed for high abundance MS1 features where the difference between conventional shotgun sequencing (Fig. 3, black dashed line) and MS2 continuation (Fig. 3, blue dashed line) is much less dominant. This reflects the MS2 selection bias towards high abundance peptides, which effectively leads to a compressed dynamic range of identified peptides.

3.3 Unsupervised MS1 feature profiling

In many proteomic experiments protein concentration changes need to be measured across different samples of complex biological background. Applications of this type of experiment span a wide range from the reconstruction of protein profiles across protein fractionations to the analysis of protein expression in time course studies. We used *SuperHirn* to extract protein profiles from the *MasterMap* of the profiling dataset, where only those profiles were considered, which were detected in at least four of the six runs. Since the expected profiles in the dilution mixtures have one or two points of low intensities, this filtering step was introduced to extract profiles or reasonable quality. The gap statistics analysis estimated that 12 clusters were required for the *K-means* clustering (see the Results and Fig. S5 in Supporting Information).

Figure 4 shows the consensus profiles of the obtained 12 *K-means* clusters together with the corresponding six theoretical dilution profiles (dashed colored lines). While six clusters were built around a constant profile with little change between the LC-MS runs or a constant profile with missing values (light gray lines), the remaining six clusters (colored lines) correlate well with their target profiles of the dilution schema (see Fig. S5 in Supporting Information illustrating the constructed *K-means* clustering profiles together with their SDs). The reproducibility of the experimentally created protein profiles demonstrate that *K-means* clustering analysis represents a valuable tool to automatically detect the main profiling trends in an LC-MS experiment in an unsupervised manner. Further examination of the constructed *K-means* clustering structure showed that all identified feature of the six dilution proteins were correctly grouped into the cluster of their corresponding target profile (see Excel file "featureClusterList.xls" in Supporting Information for a detailed listing of all MS1 features and their cluster belonging).

3.4 What is the best candidate protein?

We developed a statistical method to assign high correlation protein profiles to a given target profile. While it would be possible to simply rank protein profiles by their profile score to a target profile, this strategy would require the definition of an optimal score threshold to discriminate between truly and randomly matching profiles, a value that would largely depend on the experimental conditions. In contrast, we chose a method in which a given score was transformed into a probability that the assignment of a score to a class is correct (see Section 2.10.4), a solution that has already been described for other proteomic applications [20, 36]. The statistical method is here exemplified for the target profile of the dilution schema of aldolase A. Initially, MS1 features contained in the cluster, which correlates best to the target profile of aldolase A, were assembled into deconvoluted peptides as described before (see Section 2.10.3). Figure 5 shows the distribution of profile scores obtained by correlating the target profile with the profiles of all constructed deconvoluted peptides (Fig. 5, gray histogram). A clear partition of the histogram into two subdistributions is observed; scores with low values reflect the population of high correlation profiles and the right side of the histogram defines the population of low similarity scores. The shape of these bimodal distributions

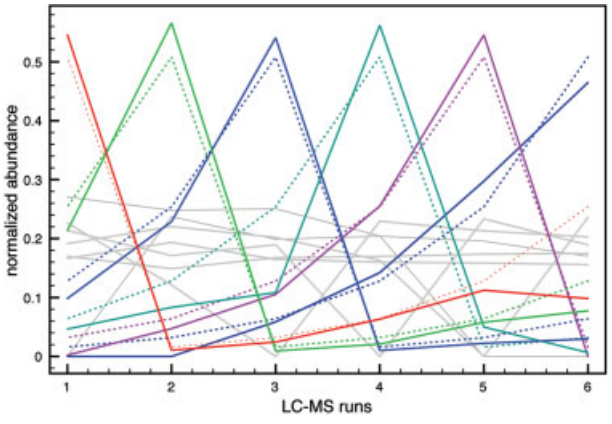

Figure 4. *K-means* clustering of MS1 feature profiles. The consensus profiles of the 12 *K-means* clusters are shown. Theoretical dilution profiles (colored dotted lines) and the best fitting cluster profiles (colored lines) are colored, while other obtained cluster profiles are depicted in gray.

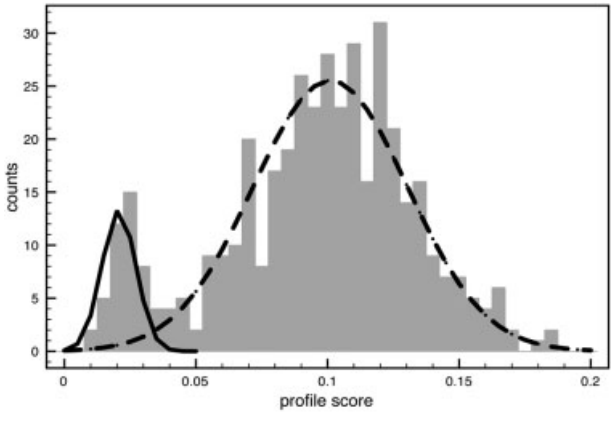

Figure 5. Evaluation of protein profile correlation to the target profile. Histogram of profile scores from charge state deconvoluted MS1 features to a given target profile (gray bars) and the fitted two-component Gaussian model of true (black line) and false profile correlations scores (black dashed line) are shown.

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

differ from experiment to experiment, and in order to adapt the calculations to the data, a two-component Gaussian model was fitted to the data (Fig. 5) and peptide profile probabilities were calculated. Deconvoluted peptides containing identified MS1 features were then assembled into proteins and profile correlation was used to remove deconvoluted peptides with low profile similarity to the protein consensus profile (see the Results and Fig. S6 in Supporting Information). Protein profile probabilities were then calculated and used to selected protein candidates (see the Results and Fig. S7 in Supporting Information).

3.5 Summary of targeted protein profiling

This statistical test was used to extract a list of high correlation protein candidates for each target profile 1–6. Candidate proteins were required to have a protein profile probability higher than 0.5. Table 2 shows the assigned target profile for every obtained protein, their average profile probabilities, the average number of identified MS1 features and in how many replicates a protein has been assigned to the correct target profile. As can be seen, for every target profile the correct standard protein was detected at a high profile probability consistently in all replicates. The reproducibility is also reflected in the number of identified MS1 features *per* protein, which does not fluctuate across replicate samples. Only for the carbonic anhydrase a lower profile probability of 0.699 with a higher SD than for the other proteins was obtained. This was caused by a low profile probability in one replicate set where the EM modeling produced overlapping Gaussian distributions. In addition, the Bovine Superoxid Dismutase (IPI00218733, target profile 2, $p = 0.499$) was detected with a good correlation to its target profile indicating that this protein is a likely contamination of the original carbonic anhydrase protein sample. However, it was identified by only one MS1 feature and not seen across all replicates.

Table 2. Summary of the targeted protein profiling experiment. For every target profile assigned protein(s) with a profile probability above a threshold of 0.5 are shown

Protein name	T^A	P^B	C^C	C^D
Myoglobin	1	1.000 ± 0.000	8.00 ± 2.16	3
Carbonic anhydrase	2	0.699 ± 0.47	6.67 ± 0.94	3
IPI00218733	2	0.499 ± 0.49	1.00 ± 0.00	2
Cytochrome *c*	3	1.000 ± 0.000	14.00 ± 0.82	3
Lysozyme	4	1.000 ± 0.000	6.33 ± 0.47	3
Alcohol dehydrogenase	5	1.000 ± 0.000	6.67 ± 0.94	3
Aldolase A	6	0.997 ± 0.004	14.33 ± 2.49	3

Averaged values and SDs are calculated for each protein over the three replicate sets of the dilution experiment. T^A, assigned target profile; P^B, protein profile probability; C^C, number of identified MS1 features *per* protein; C^D, in how many replicates the protein has been detected.

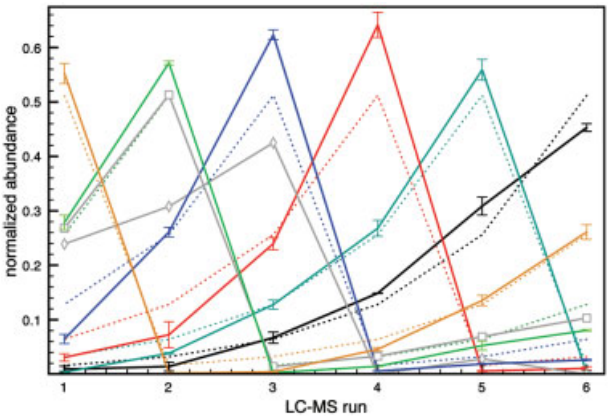

Figure 6. Averaged protein profiles of protein candidates. Consensus profiles of proteins with profile probability >0.5 are plotted together with target profiles 1–6 (dashed lines). Error bars represent standard errors over the three replicate sets. Myoglobin (orange), carbonic anhydrase (green), Cytochrome *c* (blue), lysozyme (red), alcohol dehydrogenase (violet) and aldolase A (black), and IPI00218733 (gray square).

Figure 6 shows the averaged protein profiles of all identified proteins with profile probability >0.5 (see Table 2). The small SDs of the profiles demonstrate that there is a high reproducibility across the replicates. While the target profiles correlate well with the averaged protein profiles of myoglobin, carbonic anhydrase, alcohol dehyrogenase, and aldolase A, the protein profiles of cytochrome *c* and lysozyme deviates a bit stronger from the expected abundances (Fig. 6). The profile of cytochrome *c*, for example, is very reproducible over replicates but shows at low concentrations (LC-MS runs 1 and 2) a significant deviation from the expected abundance value and due to profile normalization this results also in an increased response at the high concentrations (LC-MS run 3).

4 Discussion

The presented software *SuperHirn* unifies various functionalities for the processing of mass spectrometric data and constitutes a novel platform for the label-free quantification of multidimensional LC-MS data. The concept of LC-MS similarities is implemented into the presented multiple LC-MS alignment strategy to automatically combine LC-MS runs into a *MasterMap*, which serves then as a framework for further downstream data analysis. While LC-MS similarity scores were here utilized to construct an alignment topology, this functionality comprises more potential for data quality assessment; low quality LC-MS runs with poor similarity scores can be detected by simple clustering methods and excluded from further data analysis. The present methodology was applied to a specially designed benchmark data and demonstrated that the continuation of MS2 information across aligned MS1 features, which is feasible in high mass

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

accuracy LC-MS data, largely increases the reproducibility of identified features in LC-MS maps.

K-means clustering analysis was performed on the benchmark data and all six protein dilution trends were automatically detected. Unsupervised profiling analysis is a valuable tool for scenarios, where no *a priori* knowledge about protein concentration changes is available and data analysis is performed in discovery mode. For example, clustering analysis of protein expression in time course experiments detects groups of MS1 features, which have a similar concentration change (data not shown) without the requirement of MS2 information. These concentration changes are reflected in the profiles of the constructed *K-means* clusters and represent automatically detected target profiles for subsequent targeted peptide/protein profiling. In further downstream analysis, the detected feature groups can then be complemented by GO annotations or other genomic data to link protein expression to protein function.

Many proteomics experiments focus on specific groups of proteins that change their abundance in response to experimental perturbations in a predefined way. Using the presented benchmark dataset, we showed that the proteins could be unambiguously attributed to their dilution profiles using a probabilistic scoring method. By combining MS1 profiling analysis with targeted MS2 sequencing, non-identified MS1 features with high correlation to a target profile represent candidates for subsequent targeted MS1 features annotation efforts.

We demonstrated the methodology to map MS1 features across LC-MS runs in a multiple LC-MS alignment process. However, the alignment module of *SuperHirn* is directly applicable to a higher level meta alignment where *MasterMaps* from different experiments are combined, which offers the possibility to merge LC-MS data from different experiments and compare the aligned MS1 features of the *MasterMaps*. While the presented LC-MS data were here utilized to demonstrate the outlined profiling approach, we believe it represented in addition a valuable evaluation dataset to assess and compare the performance of different LC-MS quantification programs.

The authors thank Ralph Schiess and Matthias Gstaiger (IMSB, ETH, Zurich, Switzerland) for helpful discussions. Ning Zhang and Vagisha Sharma (ISB, Seattle, USA), Damon May and Martin Mcintosh (Fred Hutchinson Cancer Research Center, Seattle, USA), Matej Oresic (Technical Research Centre of Finland, Tietotie, Finland), Mikko Katajamaa (Turku Centre for Biotechnology, Turku, Finland), Eva Lange (Algorithmic Bioinformatics, Free University Berlin, Germany), Marc Sturm (Center for Bioinformatics, Eberhard Karls University Tübingen, Germany), Jake Jaffe (The Broad Institute of Harvard and the Massachusetts Institute of Technology, Cambridge, USA), Xiaojun Li (Homestead Clinical Corporation, Seattle, USA) and Paul Benton and Gary Siuzdak (The Scripps Research Institute, La Jolla, USA) are acknowledged for their collaboration in the software comparison. Oliver Rinner was supported by the Roche Research Foundation and Deutsche Forschungsgemeinschaft (DFG). Bernd Bodenmiller is the recipient of the Boehringer Ingelheim Fonds fellowship. This project has been funded in part by ETH Zurich, and with Federal (US) funds from the National Heart, Lung, and Blood Institute, National Institute of Health, under contract No. N01-HV28179.

5 References

[1] Aebersold, R., Mann, M., *Nature* 2003, *422*, 198–207.

[2] Li, X. J., Yi, E. C., Kemp, C. J., Zhang, H., Aebersold, R., *Mol. Cell. Proteomics* 2005, *4*, 1328–1340.

[3] Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M. *et al.*, *Mol. Cell. Proteomics* 2006, *5*, 652–670.

[4] Callister, S. J., Barry, R. C., Adkins, J. N., Johnson, E. T. *et al.*, *J. Proteome Res.* 2006, *5*, 277–286.

[5] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F. *et al.*, *Nat. Biotechnol.* 1999, *17*, 994–999.

[6] Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 1154–1169.

[7] Schmidt, A., Kellermann, J., Lottspeich, F., *Proteomics* 2005, *5*, 4–15.

[8] Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B. *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 376–386.

[9] Wang, P., Coram, M., Tang, H., Fitzgibbon, M. P. *et al.*, *Biostatistics* 2006, *8*, 357–367.

[10] Meng, F., Paweletz, C. P., Mazur, M. T., Deyanova, E. G. *et al.*, *J. Am. Soc. Mass Spectrom.* 2006, *18*, 226–233.

[11] Prakash, A., Mallick, P., Whiteaker, J., Zhang, H. *et al.*, *Mol. Cell. Proteomics* 2005, *5*, 423–432.

[12] Fischer, B., Grossmann, J., Roth, V., Gruissem, W. *et al.*, *Bioinformatics* 2006, *22*, 132–140.

[13] Jaffe, J. D., Mani, D. R., Leptos, K. C., Church, G. M. *et al.*, *Mol. Cell. Proteomics* 2006, *5*, 1927–1941.

[14] Kohlbacher, O., Reinert, K., Gropl, C., Lange, E. *et al.*, *Bioinformatics* 2007, *23*, e191–e197.

[15] Bellew, M., Coram, M., Fitzgibbon, M., Igra, M. *et al.*, *Bioinformatics* 2006, *22*, 1902–1909.

[16] Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P. *et al.*, *Nature* 2003, *426*, 570–574.

[17] Zhang, H., Li, X. J., Martin, D. B., Aebersold, R., *Nat. Biotechnol.* 2003, *21*, 660–666.

[18] Zhang, H., Aebersold, R., *Methods Mol. Biol.* 2006, *328*, 177–185.

[19] Keller, A., Eng, J., Zhang, N., Li, X. J., Aebersold, R., *Mol. Syst. Biol.* 2005, *1*, 2005.0017.

[20] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., *Anal Chem* 2002, *74*, 5383–5392.

[21] Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M. *et al.*, *Nat. Biotechnol.* 2004, *22*, 1459–1466.

[22] Gras, R., Muller, M., Gasteiger, E., Gay, S. *et al.*, *Electrophoresis* 1999, *20*, 3535–3550.

[23] Lipton, M. S., Romine, M. F., Monroe, M. E., Elias, D. A. *et al.*, *Methods Biochem. Anal.* 2006, *49*, 113–134.

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[24] Silva, J. C., Denny, R., Dorschel, C. A., Gorenstein, M. *et al.*, *Anal. Chem.* 2005, *77*, 2187–2200.

[25] Cleveland, W. S., *J. Am. Stat. Assoc.* 1979, *74*, 829–836.

[26] Saporta, G., *Probabilites, Analyse des Donnees et Statistique*, Gulf Publishing Company, Paris 1980.

[27] Phillips, A. J., *J. Biomed. Inform.* 2006, *39*, 18–33.

[28] Sokal, R. R., Michener, C. D., *Univ. KS Sci. Bull.* 1958, *28*, 1409–1438.

[29] Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M. *et al.*, *Nucleic Acids Res.* 2002, *30*, e15.

[30] Black, P. E., *Dictionary of Algorithms and Data Structures*, U.S. National Institute of Standards and Technology, 2006.

[31] MacQueen, J., *Proc. 5th Berkely Symp. Math. Prob.*, IEEE, Berkley, CA 1967, pp. 281–297.

[32] Xu, R., Wunsch, D., II, *IEEE Trans. Neural Netw.* 2005, *16*, 645–678.

[33] Tibshirani, R., Walther, G., Hastie, T., *Technical Report 208*, Deparment of Statistics, Standford University 2000.

[34] Dixon, W., *Biometrics* 1953, *9*, 74–89.

[35] Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning*, Springer-Verlag, New York Inc 2001, p. 549.

[36] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., *Anal. Chem.* 2003, *75*, 4646–4658.

[37] Montgomery, A. A., Peters, T. J., Little, P., *BMC Med. Res. Methodol.* 2003, *3*, 26.

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim