REVIEW

Comparative LC-MS: A landscape of peaks and valleys

Antoine H. P. America and Jan H. G. Cordewener

Plant Research International, Wageningen University and Research Centres, Wageningen, The Netherlands

Quantitative proteomics approaches using stable isotopes are well-known and used in many labs nowadays. More recently, high resolution quantitative approaches are reported that rely on LC-MS quantitation of peptide concentrations by comparing peak intensities between multiple runs obtained by continuous detection in MS mode. Characteristic of these comparative LC-MS procedures is that they do not rely on the use of stable isotopes; therefore the procedure is often referred to as label-free LC-MS. In order to compare at comprehensive scale peak intensity data in multiple LC-MS datasets, dedicated software is required for detection, matching and alignment of peaks. The high accuracy in quantitative determination of peptide abundancies provides an impressive level of detail. This approach also requires an experimental set-up where quantitative aspects of protein extraction and reproducible separation conditions need to be well controlled. In this paper we will provide insight in the critical parameters that affect the quality of the results and list an overview of the most recent software packages that are available for this procedure.

Received: July 15, 2007
Revised: November 1, 2007
Accepted: November 1, 2007

Keywords:
Alignment / Comparative LC-MS / Label-free / Quantitative proteomics / Software

1 Introduction

The huge complexity, enormous concentration range and dynamic nature of the proteome present great technological challenges for proteomic research. The physiological status of a cell or organism is dictated by a multitude of regulatory processes taking place at several levels. The steady state concentration of each protein, the balance of its synthesis and turn-over rate, protein cleavages and other modifications, interactions with a network of binding partners, localisation within the cell; all affect the effective activity of each protein and, in total, the status of the organism.

To gain insight of this complex regulatory network and to identify those components that have major influence is the goal of many proteomics studies. Biomarker identification is aiming at the selection of those cellular components that are consistently correlated with a particular physiological status like development or disease [1–4]. Biomarker discovery essentially comes down to detect consistent differences within a large number of samples. The technological challenge in biomarker discovery is to develop an efficient method for detailed and reproducible detection of as many components as possible, preferable in a quantitative manner.

LC-MS has evolved into a powerful procedure for detailed identification as well as quantification of complex proteomic samples. As opposed to 2-DE, where samples are separated at the protein level, LC-MS analysis is mostly performed at the peptide level [5, 6]. Protein extracts are treated with protease (trypsin) in order to convert the sample into a complex mixture of peptides, which can be analysed in detail with MS. For analysis of highly complex protein mixtures a multi-dimensional separation is often applied, where the first dimension separation can be performed either at the protein level (before trypsin treatment) [7–9] or at the peptide level [10–12].

While initially most proteomics research was limited to the detection of qualitative differences in protein composition between samples, nowadays the accuracy of modern MS analysis enables detailed detection of quantitative differences even down to relative ratios of two or less [13–18]. This review

Correspondence: Dr. Twan America, Plant Research International, Wageningen University and Research Centres, PO Box 16, 6700AA Wageningen, The Netherlands
E-mail: twan.america@wur.nl
Fax: +31-317-418-094

Abbreviations: AMT, accurate mass-time; **AMRT,** accurate mass retention time pair; **PCA,** principal components analysis; **PMT**, peptide mass-time; **Rt,** retention time

is intended to provide an overview of recent approaches in quantitative LC-MS analysis that are based on a quantitative comparison of peak patterns between samples, mostly without the use of stable-isotope labels.

2 Quantitative proteomics

Most of the quantitative proteomics approaches by MS utilise isotopic labels as a reference for either relative or even absolute quantification. Stable isotopes can be introduced *in vivo,* by feeding cells or an entire organism with a medium enriched with stable isotopes [19]. Alternatively, the isotope can be introduced into the proteins after extraction from the sample, using a covalent coupling reagent that contains either the natural or the heavy form of the isotope [20–22]. Isotope labeling enables multiplexing. By mixing the differently labeled samples before analysis (or even before extraction) experimental procedures can be performed on the mixture of samples. Quantification of changes in protein concentration is then performed by comparing the signal intensities of peptide ions containing the stable isotope *versus* the natural compound [23–25]. Any errors in the sample handling, extraction, separation and detection equally affect both samples at the same rate and thus are compensated by the parallel analysis. Quantitative proteomics approaches using stable isotopes are well-known and used in many labs nowadays.

Many quantification methods using stable isotope procedures require MS/MS detection of peptides in order to identify the sequence and by that the corresponding isotope-labeled mass peak. This is especially so for the *in vivo* labeled samples, where the mass shift of the peptide is dependent on the number of incorporated isotope molecules, which is dictated by the peptide sequence [26–31].

The detection coverage of a complex peptide mixture using LC-MS/MS approaches is limited, however. The sequential procedure of selecting peptides for fragmentation, one by one, while they are eluting from an HPLC column, confers a maximum to the number of peptides that can be selected in a certain amount of time. Depending on the acquisition speed of the mass-spectrometer in use, one to five fragmentation spectra can be collected *per* second. Given the complex nature of the proteome, a tryptic digest of a protein extract (coming from an affinity enriched sample up to a total tissue extract) may contain from a few hundred up to some hundreds of thousands of peptides. The detection coverage is at stake, even more so in the latter case.

Chromatographic separation upfront to the MS detection step has no capacity to separate such complex peptide mixtures at baseline level. Overlap of multiple eluting peptide is the result. In highly complex peptide mixtures several tenfolds of different peptides may be co-eluting at any moment in time. Multidimensional separation procedures help to improve resolution of detection, however, at the cost of the number of samples that can be analysed in a practical way [8–12, 32, 33].

Data dependent MS/MS detection can therefore result in undersampling of complex peptide mixtures [34]. In addition, depending on the software and parameters used for data dependent MS/MS acquisition, the set of selected peptides is actually biased to the most abundant peptides in the sample. Liu *et al*. presented and tested a statistical model for frequency of peptide selection for MS/MS [34]. They demonstrated that of the abundant classes of proteins 80 to 98% were detected in a triplicate 2-D LC-LC-MS/MS experiment. However of the lower abundance classes of proteins (which are the majority) only 10 to 25% were detected at all. In the same study it was shown that the frequency of MS/MS selection could actually be used as a quantitative measure for the relative peptide abundance. This principle led to an alternative procedure for label-free quantitation, called spectral counting. A normalised spectral abundance factor was used as measure for relative peptide quantitation [17, 34–37].

On the other hand, quantitation procedures that rely on MS (in contrast to MS/MS) detection are much less constrained by the peptide selection process, since peak resolution in MS mode is actually very high. Hundreds of co-eluting masses can be detected *per* single MS scan. As such, tens of thousands of peptides can be detected in a single LC-MS run. On this principle, quantitative approaches are developed that rely on LC-MS quantitation of peptide concentrations by comparing peak intensities between multiple runs obtained by continuous detection in MS mode (see Table 1) [13–16, 33, 38–53]. Characteristic of these comparative LC-MS procedures is that they do not rely on MS/MS or the use of stable isotopes, though in many cases would be compatible with the use of stable isotopes.

Old *et al*. performed a detailed comparison of the spectral counting *versus* the peak intensity procedure for protein quantification in a multidimensional LC-MS/MS setup [17]. They demonstrated that peak intensity measurements displayed more accurate estimates of protein ratios. The sensitivity of peak detection was limited in their case to those peptides selected for MS/MS. In more recent comparative LC-MS procedures all peaks detected above noise level in MS mode can be quantified as described in this review.

In this paper we will describe in more detail the procedures that are applied in the approach of (label-free) comparative LC-MS using peak intensity for peptide quantitation. We will provide insight in the criticical parameters that affect the quality of the results and list an overview of the most recent software packages that are available for this procedure.

3 Comparative LC-MS

Comparative LC-MS approaches rely on the observation that the peak intensity (or better: peak volume as detected in LC-MS mode) in most cases is proportional to the concentration of the peptide in the sample [13, 15, 54, 55]. So, determining the peak volume for each mass peak and comparing these

volume data between multiple LC-MS runs of different samples will provide a comprehensive quantitative overview of thousands of peptide concentrations between the samples. From this overview, a selection list of differential peptides can be produced for subsequent targeted fragmentation by LC-MS/MS in order to obtain sequence information of the selected peptides.

Though this principle may sound very straightforward, in reality this approach imposes some practical constraints in experimental approach as well as technical constraints to the instrumentation and the processing software, also reviewed in [56, 57]. A detailed and accurate quantitation of a maximum number of detected mass peaks requires preferably a high resolution mass spectrometer and a highly reproducible (nano)HPLC separation procedure. Experimental drifts in *m/z* and retention time (Rt) significantly complicate the direct comparison of multiple LC-MS datasets. In order to compare at comprehensive scale peak intensity data between samples dedicated software is required. Several steps in data processing are performed: Peaks have to be distinguished from background noise and from neighbouring peaks (peak detection). Peak integration produces peak volume data. Deconvolution takes care of charge detection and detection (removal) of isotope patterns. And most important, chromatographic alignment of elution profiles is required in order to correctly match the corresponding mass peaks between multiple LC-MS runs. Finally normalisation procedures enable a more accurate matching and quantitation between multiple samples. For correct quantitative analysis, correct peak detection and peak matching is crucial. Therefore the quality of the alignment software is a key parameter in the comparative LC-MS procedure. This will be discussed in more detail below in this article.

A typical workflow of the analytical procedure for comparative LC-MS is depicted in Fig.1a. Protein extracts from different samples are digested by protease (mostly trypsin) to prepare complex peptide mixtures. An aliquot of the peptide mixture *per* sample is injected into the LC-MS system. The remainder of the sample digest is used for replicate injections and for future (targeted) LC-MS/MS analyses. During LC-MS acquisition the mass spectrometer is set, in most cases, to acquire MS spectra only. A high frequency of MS spectra is required in order to obtain enough chromatographic resolution for the analysis software to be able to perform correct peak detection and peak integration. After peak detection and alignment is performed by the LC-MS analysis software, a quantitative comparison between samples will enable the selection of those peaks that display differential behaviour between samples. As data are collected in MS mode only, the identity (sequence) of the selected peaks is yet unknown at this stage. For this, another aliquot of the sample is injected in the LC-MS system, where the mass spectrometer now is instructed to selectively collect MS/MS data from these differential masses which are enlisted in a so-called include list. In this approach the quantitative information and the identification information are collected in separate LC-MS acquisitions, which may complicate the data analysis. On the other hand, the advantage is that MS/MS

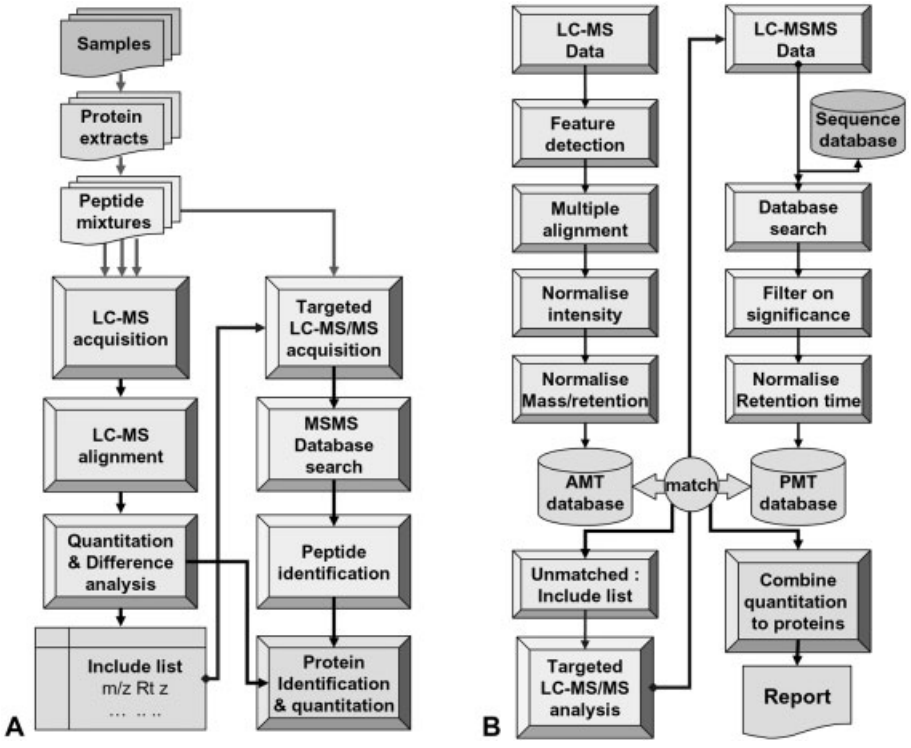


Figure 1. (A) Typical workflow of the analytical procedure for comparative LC-MS is given. Depending on the MS hardware situation, MS/MS spectra can be collected in parallel or sequentially to the LC-MS acquisition. Targeted LC-MS/MS selection is based on include list information composed after quantitative difference analysis. (B) Overview of the data processing workflow. Not all processing steps are provided in some software packages. The order of processing also differs between packages.

collection is selectively triggered for those peptides that are actually of interest because of their differential occurrence. In recent approaches, concurrent acquisition of MS and MS/MS data is possible, such as in MS systems with parallel acquisition (*i.e.* FT-ICR or FT-OrbiTrap), where the MS/MS spectra are collected in the ion trap, in parallel to the MS acquisition in the FT detector. The alternate scanning approach, as developed by Waters Corporation, where alternating MS scans with low and high collision are acquired, provides another means of collecting intensity as well as identification data in a single LC-MS run [16, 58, 59]. Whereas the MS/MS information collected during the LC-MS run will not contain complete sequence information for all differential peptides, the limited amount of MS/MS information can also be used as landmark information in the alignment procedure (see section 4.5 below) [60].

As comparative LC-MS provides highly detailed, quantitative results, statistical analysis can be performed on the data. This requires replicate injections *per* sample. For this, multiple runs of LC-MS should be performed under near identical conditions. Drifts in Rt should be minimised (same column, same gradient, preferably temperature controlled). The use of a reference protein digest as internal standard spiked into each sample will improve on the accuracy by intensity normalisation (and in some cases peak matching) between multiple samples [16, 58, 59]. Internal reference peptides can also be used for standardisation of mass and Rt, in order to compare datasets that were collected at a different moment (under different separation conditions) see section 4.6 below.

Comparative LC-MS is a relatively recent procedure. For good quality results an optimal combination of practical procedure and good quality data processing is required. For optimal data processing it is important to understand the parameters that influence the results of data extraction. In recent years several software packages have been presented. An overview of LC-MS alignment software for proteomics applications (as currently known to the author) is presented in Table 1. These software packages use different algorithms for data processing. Not all of these have been described in all detail. In the remainder of this article we will describe the different steps in data processing, comparing the different algorithms.

4 Software

A first classification of the software can be made with respect to its availability; open-source, commercial and custom (in-house) packages. Quite a number of open source (or freely available (academic) packages) have been described in literature. Many of these packages are still in a (relatively early) stage of development. Other packages, like MetAlign, MSight, MsInspect, PEPPeR, SuperHirn and VIPER are fully operational, such that analysis of complex peptide LC-MS datasets have been reported in literature and that the software has more or less clear (and easy to install) user interface components. Recently, a number of commercial packages have reached the market. Whereas also these packages are open for improvement, user interface and data processing algorithms appear to be more evolved. More attention has been paid to make the programme accessible to novel users. A small number of packages have been described in literature that are custom built in a commercial setting, and are not (yet?) available to the MS users' public (see Table 1).

4.1 Data format

Discrimination can be made as to the input format that is used by the software. The commercial packages SIEVE and PLGS Expression, developed by MS hardware providers, are able to directly process the raw file format as produced by the MS hardware. This is an advantage, as no data conversion processing is required. At the same time this limits the use of the software to one particular "brand" of MS data. On the other hand, open source packages in most cases use the mzXML format as a generic data format, for which several converter modules are available [79, 80]. This broadens the applicability of the software. However, a disadvantage of mzXML format is the considerable increase in data size (*circa* 3–5 fold) and processing time required for data conversion. As a single LC-MS raw data file from a 90 min gradient elution can be *circa* 1 Gb in size, its mzXML counterpart can be up to 5 Gb in size, after *circa* 1 h of conversion calculation. A medium sized LC-MS experiment, containing for example 20 LC-MS runs, then will result in 20 Gb (raw) plus 100 Gb additional (mzXML) data, even prior to any data analysis. This not only puts constraints on the data storage capacity, it also puts strong requirements on memory managenement for the processing PC, in order to be able to process these large size data files. Most recently, the Proteomics Standards Initiative is developing a new format intended to replace the mzXML and mzData formats [81]. The new format will be named mzML, and should allow compressed file formats for compressed peaklists. However, data export and import filters of any existing programs need be adapted in order to make this format usable in practice.

4.2 Peak detection

A flow scheme of the different data processing steps is represented in Fig. 1b. In most packages peak detection and feature extraction is the first step in data processing. A feature is defined as a collection of m/z peaks that are derived from the same molecular ion, as result of ^{13}C isotope distribution and/or multiple charge states distributed over multiple consecutive MS scans as result of chromatographic elution. This step involves noise filtering (smoothing), background subtraction, peak detection and grouping of m/z peaks over multiple consecutive MS scans for each detected peak. This step results in a large reduction in data size, as the raw data format is converted from a continuous data format

Table 1. Overview of LC-MS alignment software for proteomics solutions

Software name	Supplier / author	Database/ environment	availability	Functionality	website	reference
PLGS IdentityE Expression Informatics	Waters Corp	PLGS	commercial	f, h,i,b,a,r,s,l	http://www.waters.com/	[15, 16, 58, 59]
SIEVE	Thermo Scientific	BioWorks	commercial	p,h,v,b,a,r,s,l	http://www.thermo.com/	
DeCyderMS	GE Healthcare		commercial	f,h,i,b,v,a,r,s,l	http://www.gelifesciences.com/	[50]
Rosetta Elucidator	Rosetta Biosoft-ware		commercial	f,h,i,b,v,a,r,s,l	http://www.rosettabio.com/products/elucidator/default.htm	
MS-Xelerator	MsMetrix		commercial	f,l,i,b,a,r,s,	http://www.msmetrix.com	
MassView	SurroMed		custom	f,l,i,b,a,r,s,l		[61–63]
MetAlign	WUR		free for acad.	p,l,b,a,s	www.metalign.wur.nl	[33]
MzMine	VTT Finland		open source	f,h,v,a,r	http://mzmine.sourceforge.net/index.shtml	[42, 43]
MSight	SIB		open source	f,h,i,v,(a)	http://www.expasy.org/MSight/	[64]
MS Inspect	CPL (Fhcrc)	CPAS	open source	f,h,v,a,r (l,d)	http://proteomics.fhcrc.org/CPL/msinspect.html	[41, 65]
SpecArray	ISB /SPC	TPP	open source	f,h,i,v,a,r,s	http://tools.proteomecenter.org/SpecArray.php	[49]
PePPer	BROAD MIT	Genepattern	open source	h,a,r,s,l	http://www.broad.mit.edu/cancer/software/genepattern/desc/proteomics.html	[60, 66]
VIPER	PNNL	PRISM	open source	f,h,i,b,v,a,r,s,l,d	http://ncrr.pnl.gov/software	[57, 67]
OpenMS	Berlin Saarland Tubingen Univ.	TOPP	open source	(f,h,i,b,v,a,r,s,l,d)	www.openMS.de	[68–70]
SuperHirn	IMSB @ETH	Corra	open source	f,i,b,v,a,r,s	http://tools.proteomecenter.org/SuperHirn.php	[53]
CPM (continous profile models)	Listgarten/Emili	MatLab	free for acad.	l,a	http://www.cs.toronto.edu/~jenn/CPM/	[52]
Xalign	Purdue Univ	Xmass	upon request	(f,h),i,a,s	zhang100@purdue.edu	[71]
Fischer et.al.	ETH		not described	h,a	http://people.inf.ethz.ch/befische/	[46]
CRAWDAD	Washington Univ		upon request	f,l,i,a,r,s,l,d	http://proteome.gs.washington.edu/software/crawdad	
CHAMS	Inst Pasteur, Paris		web server	h,a,s	http://www.pasteur.fr/recherche/unites/Biolsys/chams/index.htm	[51, 72]
OBI-WARP	Univ. Texas		open source	a,r,l	http://bioinformatics.icmb.utexas.edu/obi-warp/	[73]
LCMSWARP	PNNL	PRISM	open source	h,a	http://ncrr.pnl.gov/software	[74]
LCMS2D	Albert Einstein College of Medicine				http://www.bioc.aecom.yu.edu/labs/angellab/	[75, 76]
PETAL	CPL (Fhcrc)	CPAS	open source	a	http://peiwang.fhcrc.org/research-project.html	[77, 78]

p/f: peak/feature detection; h/l: high/low resolution; i: de-isotoping; b: batch processing; v: LCMS 2-D visualization; a: alignment; r: result visualization; s: statistical analysis; l: link MS to MS/MS; d: results database

to a list of discrete identified features. As this is a calculation intensive step, batch processing of multiple LC-MS datasets is a clear advantage (it may require more then an hour *per* LC-MS run, so overnight processing for multiple LC-MS runs). Several algorithms for peak detection have been reported [42, 62, 69–71, 75, 76, 82–85]. A detailed discussion of the different algorithms is presented in [56]. The different approaches have effect on the consistency and quality of the resulting peak list. The best peak detection results are reported from algorithms that take both the *m/z* and the time dimension into account [76]. Peak detection is in most cases performed on a file *per* file basis (*i.e. per* individual LC-MS run). However, consistency of peak detection over multiple LC-MS datasets is an important aspect, often neglected. For instance, not all *m/z* peaks display equal peak width in an LC-MS chromatogram. Where early eluting and medium intensity peaks can be most narrow, peak width will increase at later elution times, but also for high intensity peaks, which often display prominent tailing. If tailing peaks are detected as multiple consecutive peaks, there is a risk of errors in peak matching during later data processing, resulting in a decrease of quantitative data quality in the final results table. Care should be taken to minimise peak width as much as possible during LC-MS acquisition (high resolution nanoLC conditions, high efficiency LC columns, minimal dead volume connections, no column saturation), as this will increase the chromatographic resolution, and as such the quality of peak detection. A procedure for detection and clustering of peak repeats is described by de Groot *et al.* [86]. In some cases peak detection (or framing) is performed on the complete (aligned LC-MS) dataset (as in SIEVE). This improves the consistency of data, minimising the occurance

of missing datapoints. A drawback of the SIEVE algorithm, on the other hand, is that it uses a fixed sized frame for peak detection, which does not seem to be apropriate for quantifying peaks with varying peak widths.

4.3 LC-MS visualisation

Although not absolutely necessary for comparative LC-MS analysis, visualisation of the LC-MS dataset (raw or mzXML data) in combination with the detected LC features can provide a detailed insight in the quality of the LC-MS separation as well as of the quality of the feature detection (see Fig. 2). Complications like extensive peak tailing (chromatographic saturation), strong ion suppression effects, or unequal distribution of peak separation and peak overlap (due to non-optimal elution gradient profile) are evident in an LC-MS visualisation. The software packages MsInspect, MSight, VIPER, SpecArray and DecyderMS provide full-featured visualisation tools for zoomable inspection of the LC-MS data. Also the quality of feature detection can be evaluated in an LC-MS visualisation. Completeness, uniqueness, and resolution of feature detection and charge detection can be scrutinised, which may help in optimising the peak detection parameter settings. However, the alignment capacity of MsInspect, Msight and SpecArray is currently limited, such that these packages enable comparison of only a few LC-MS runs.

On the other hand, the high complexity of multiple protein digests constrains the visual inspection mainly to a global (and subjective) quality control. The sheer size of the multiple LC-MS data will not allow detailed inspection or correction of all detected features. The user needs to trust the detection algorithm, once the optimal peak detection parameters have been determined, A more robust quality evaluation of the total data analysis procedure should be performed by statistical evaluation of the results [56] (see Section 4.4). Piening *et al.* have described a quality control metrics approach for assessing the accuracy of feature detection in the absence of completely annotated data sets [87].

4.4 Feature filtering

Some packages provide feature filtering options, in order to remove low quality features from a dataset, before entering the alignment procedure. Parameters like minimal peak intensity, quality of isotope distribution, consistency of charge detection, chromatographic elution shape can be used to remove low quality detected features, which would complicate (and possibly disturb) a correct alignment of multiple LC-MS feature lists. Again, visualisation may help in selecting filter settings. MsInspect provides a sorted heatmap visualisation of all isotope distributions classified by

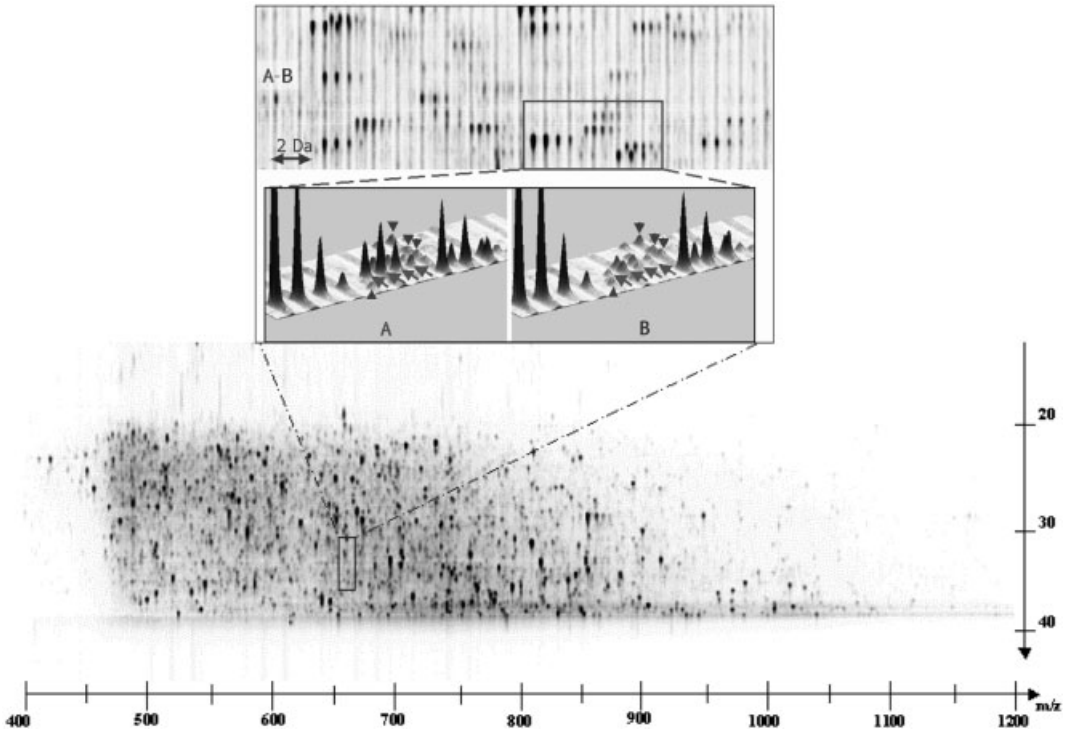


Figure 2. An example is given of LC-MS visualisation using the MSight program in this case. LC-MS peak patterns can be visualised in 2-D or 3-D views, possibly in two colour overlay. A 2-D visualisation of peak pattern in m/z and Rt dimension provides a good overview of separation quality. Zoom view provides information about resolution. (source: MSight users manual, with permission from the Proteome Informatics Group, Swiss Institute of Bioinformatics) [64]

charge state (see Fig. 3). Sorting, based on ion distribution score provides a (subjective) clue on the ion distribution threshold which separates valid from invalid charge state detections [41]

4.5 Peak alignment

Alignment of multiple LC-MS datasets is a non-trivial operation. The challenge is to find for each detected peak of one dataset a matching peak in the other dataset. Complications that can be encountered are: i) the exact *m/z* and Rt of the corresponding peak may differ slightly due to technical drift in MS and LC hardware, ii) the before mentioned drifts will not be equal for all detected peaks (especially Rt drift appears to be non-linear over a complete elution trajectory), iii) a single peak of the corresponding peptide may be detected as multiple peaks in the other LC-MS dataset (due to peak tailing). iv) other *m/z* features with a nearly similar *m/z* and Rt value but originating from a different peptide may exist in a highly complex peptide mixture v) the corresponding peak may actually not be present, simply due to absence of that particular peptide in the other sample, or may not be detected due to low intensity. Given the enormous complexity of high resolution LC-MS datasets, a perfect detailed alignment of all features seems a non-realistic goal. Several alignment algorithms have been described. Many algorithms use mass (or *m/z* and charge), Rt and optionally intensity information of detected LC-MS features in order to find an optimal matching between two datasets. Algorithms like correlation optimised warping [73, 74, 88], vectorised peaks [62], (semi-) supervised alignment using non-linear regression methods [46] or Hidden Markov Models [52], statistical alignment [77] or clustering [15, 53, 70] have been described. Clearly these alignment procedures strongly benefit from high mass accuracy and stable Rts, especially when aligning LC-MS data from highly complex peptide mixtures [57]. The more the resolution is in *m/z* detection as well as chromatographic separation, the higher the peak capacity is in both dimensions. Many of the recent alignment algorithms (PLGS, SIEVE, VIPER, Pepper, MsInspect, Msight) depend on high resolution *m/z* data. Other algorithms (MetAlign, CPM, Crawdad, MsMetrix) use data binning, in which selected ion chromatograms are extracted, often with 1 Da width. The advantage of the latter approach is that it is also appli-

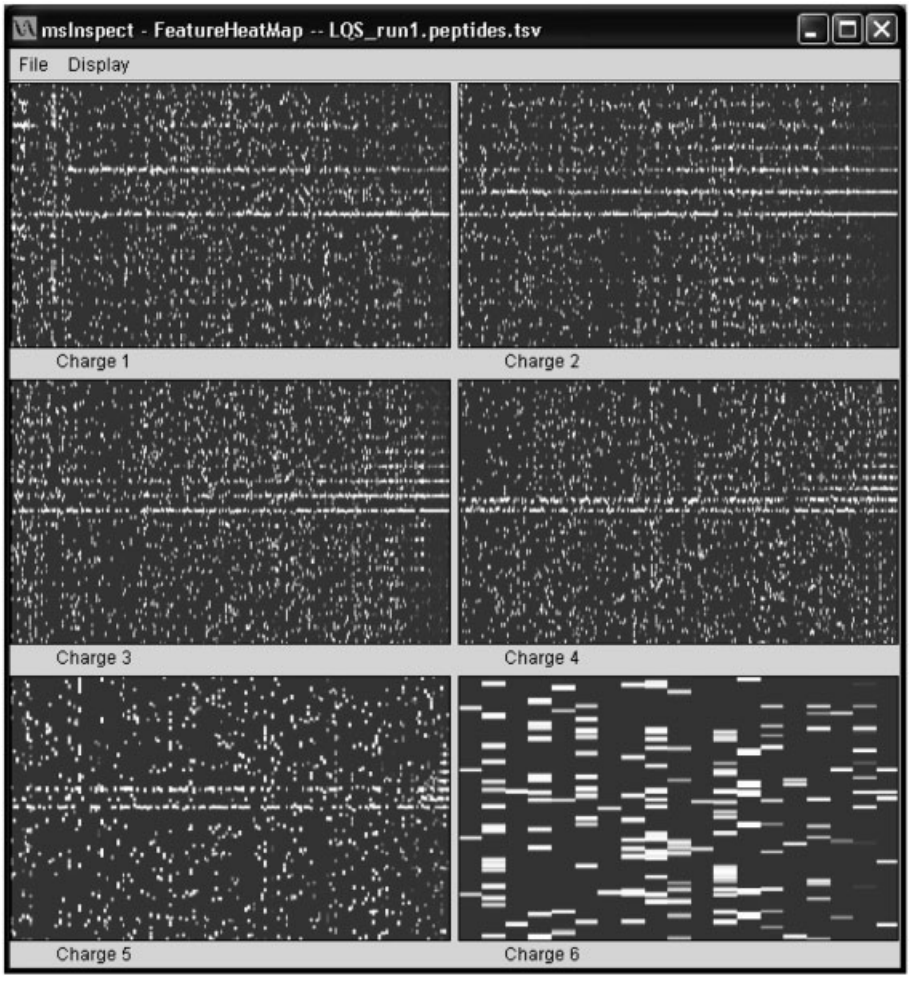


Figure 3. Heat map visualisation of isotope distributions over all detected features separated by charge state, as provided by MsInspect [41]. The isotope profiles of all detected features are displayed, vertically centered at the ^{12}C monoisotopic peak and horizontally sorted on the basis of their ion distribution score (a quality parameter for fitness of the isotope model). A feature filtering threshold can be manually set based on visual inspection of the isotope distribution.

cable to lower resolution MS data such as from ion trap instruments. However, this is at the expense of peak capacity and the risk of peak misalignment. On the other hand, surprisingly large tolerance to chromatographic drift has been demonstrated by some of these algorithms, see Fig. 4 [33, 62].

The before mentioned procedures fully rely on LC-MS-only data. This approach strongly relies on similarity between LC-MS datasets. Large deviations in peak pattern due to either changes in sample composition or drifts in Rt or *m/z*-value of peptides severely complicate the alignment process. Alignment of LC-MS data from different fractions in a multidimensional LC-MS separation is very complicated, if not impossible in this approach. Some programmes first perform a similarity clustering of LC-MS data directly based on the signal level, actually before peak detection and alignment [51]. In this way the most similar datasets are aligned to each other first [53]. In a very recent approach peptide elements (features) in combination with raw data are aligned across all LC-MS runs simultaneous using statistical methods, without time-warping. This improves symmetry within datasets and prevents missing datapoints [52, 77].

Alternative approaches have been described that combine information from LC-MS with MS/MS information. These approaches rely on data acquisition modes with fast alternate switching between MS and MS/MS mode, such as possible with modern ion trap and Q-TOF instruments, or (semi)parallel acquisition, as possible with iontrap-FT-MS or iontrap-Orbitrap instruments. In these approaches the MS/MS spectra with confidently identified peptide sequences provide a time based framework on which the LC-MS(/MS) data can be roughly aligned. A (more detailed) refinement of the alignment is then subsequently performed on the high density LC-MS datasets using the MS/MS alignment as a starting point [46, 60, 73]. This procedure enables also alignment of LC-MS datasets that display large difference, such as subsequent fractions from a multidimensional separation. The data analysis approach developed in the group of Smith provides a combination of a peptide mass-time (PMT) database, loaded with a large collection of high confidence MS/MS peptide identifications coupled to a quantitative accurate mass-time (AMT) alignment approach which is then matched to the PMT database in order to link quantification information to previously identified peptides [40, 45, 57, 67, 89](see Fig. 1b). For the matching of these databases Rt normalisation is essential (see Section 4.6). Also in this approach, LC-MS features for which differential expression has been quantified, but no identification exists, due to absence in the PMT database, will be listed in an "attention" or "include" list (listing *m/z* and Rt *per* feature) for targeted MS/MS collection in a follow-up LC-MS/MS acquisition. The advantage of such a database matching approach is that information collected from multiple experiments can be linked. A PMT collection is build for a particular organism or tissue, using multiple high frequency MS/MS acquisition experiments, acquired during or separately from the quantitative LC-MS runs. The coverage of the PMT data grows while more MS/MS runs are acquired. The AMT approach then is performed on multiple samples for quantitative analysis preferably using high resolution TOF or FTMS. Of major importance here is a high accuracy of mass and (normalised) Rt for both types of acquisitions, as these two parameters are the only link between the two databases.

An interesting (alternative) approach of parallel fragmentation and quantification analysis has been reported by Silva *et al.* [16] and Vissers *et al.* [59], making use of the alternate scanning approach (Identity^{E}) developed by Waters Corporation. In this approach, the QTOF mass spectrometer alternately switches between a low collision and (ramped) high collision energy in the collision cell in front of the TOF MS detector. Where the low energy scan provides accurate *m/z,* charge, Rt and peak intensity information, the high energy scan provides fragmentation information from nearly all co-eluting peptides, as there is no *m/z* selection in the 1st quadrupole in front of the collision cell. The fragmentation

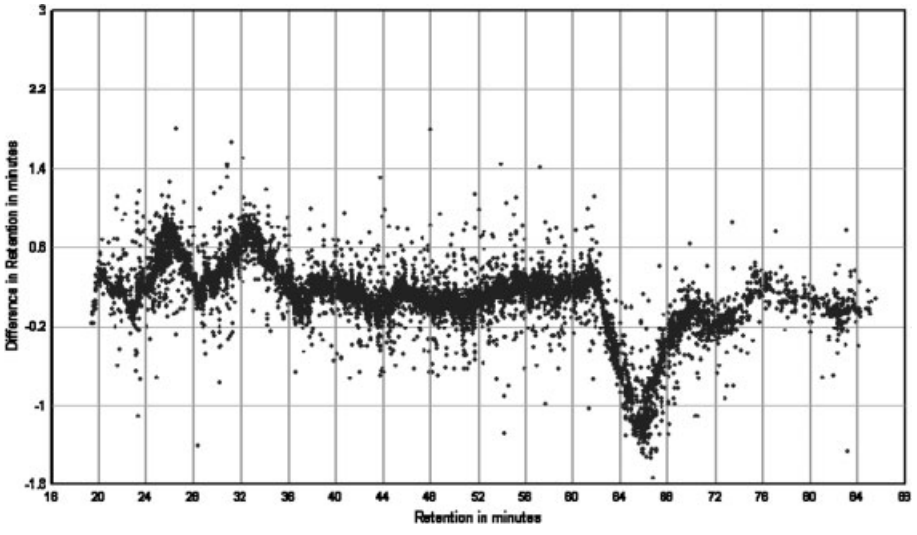


Figure 4. Rt difference plot. The difference in Rt is plotted for each peak that is matched with a peak in the reference LC-MS dataset. As many different peaks per timepoint are matched with similar retention drift a narrow band distribution is observed. Outlier datapoints are due to mismatching or error in peak detection. Retention time drift is clearly a non-linear phenomenon. Figure reproduced from [33].

information is used for a high coverage sequence identification of co-eluting peptides [59], but it can also be used as additional information for alignment between different LC-MS datasets. The fragmentation information provides additional clues as to the peptide identity, such that incorrect alignment of similar *m/z* - Rt pairs between different runs can be minimised.

The result of an alignment algorithm is the production of a match table, also called accurate mass Rt pair (AMRT) table (PLGS, VIPER) or peptide array (MsInspect, SpecArray). In such a table feature parameters like *m/z*, charge, Mw, Rt, intensity, volume and possibly peptide sequence information, *etc.* are listed in columns *per* LC-MS experiment. Each row then contains a set of features (peptides) that are matched between these LC-MS experiments. If a feature is not detected or not matched in a particular LC-MS dataset, this will result in missing data fields in the AMRT table. Within replicate LC-MS runs, the number of missing data fields should be minimal. In other words: the replicate number (=occurrence/replicate) should be maximal as, in principle, peaks should be detected and matched for each of the replicate injections [77, 86].

For many software packages this is the endpoint of the analysis. The table can be exported as delimited text file to be analysed further in other data analysis packages, like Excel or other statistics software. Some packages provide options for quality evaluation and optimisation, as mentioned below. A software package like R [90] (open source) can be used for statistics analysis. Visualisation software for multivariate data analysis like Spotfire DecisionSite [91], GeneMath [92] or other micro-array data analysis packages, like SAM [93, 94] or TM4MeV [95–97], can be very helpful in evaluating global data quality as well as selecting the differential peptide features (see Section 4.7).

4.6 Normalisation

The result of peak detection and alignment is a large table of peak intensity values coupled to (detected) mass, charge and Rt values. Advanced datafiltering and processing procedures are required to correct, refine and filter the data in order to extract meaningful information. Normalisation and transformation of quantitative data is essential for correction of global errors in data quality. Global drifts occur at various levels during and between multiple MS acquisitions. Normalisation procedures can correct for these drifts, in order to improve the accuracy of quantitative results and to standardise the results, such that matching and comparison between multiple experiments is improved. Normalisation can be applied to mass (*m/z* or Mw values) (recalibration), Rt (Rt normalisation) and/or peptide abundance (peak intensity or volume). Normalisation of mass is in some cases performed already in the alignment algorithm, as calibration to an internal reference (*e.g.* the lock-spray calibration in Waters Corporation PLGS). In other cases this could be performed afterwards. Linear transformations *per* LCMS dataset can correct for a global calibration error [65]. However, more fine resolved calibration functions can be derived from the mass errors calculated between experimental and theoretical mass for those peaks for which sequence information has been collected (in case MS/MS or MS^E data have been acquired).

Rt normalisation, or standardisation, is a more dramatic correction, as the Rt is dependent on many factors (*e.g.* column condition, temperature, eluent gradient and pump condition). In the implementation of the PMT and AMT database system the observed retention (or elution) time is regressed to a predicted normalised elution time based on heuristic models predicting the normalised elution time from the peptide sequence identified by MS/MS [98]. Prediction of peptide Rt in relation to their amino acid sequence has been modeled on the basis of a collection of identified MS/MS spectra [99, 100]. More recently, a support vector regression algorithm was presented which can be applied on local LC-MS/MS datasets in order to train a Rt model based on experimental data obtained from the same system as used for quantitative LC-MS acquisition [101].

Normalisation of peptide abundance data is probably the most essential for improvement of the quantitative accuracy of the experiment. Abundance normalisation will correct for bias due to errors in sample size, possibly sample carry-over and drifts in ionisation and detector efficiencies. Different procedures for normalisation can be applied. Normalisation values can be calculated on the basis of a global distribution for all detected features (like sum, average or median of all intensities *per* run), or calculated from a specific sub-set of features, for instance from a spiked protein that is used as internal standard [16, 58, 59], or a set of "household' proteins. In most cases a global correction factor for the complete LC-MS dataset is applied, though a local correction factor for intensity maybe more applicable just as for mass and Rt normalisation [52, 56]. A detailed discussion on normalisation and statistical evaluation of comparative LC-MS data is presented in [56] and [102].

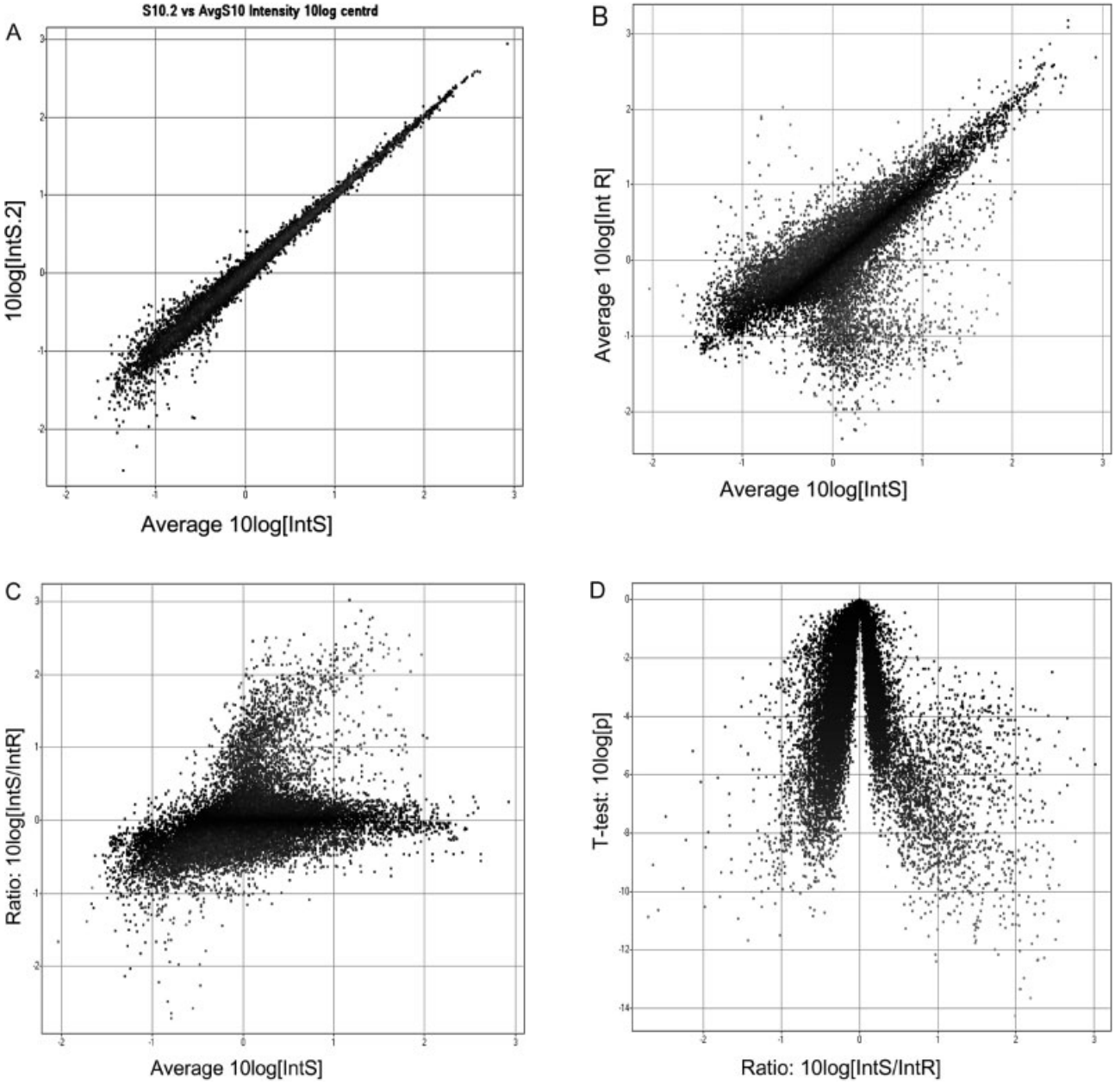
Callister *et al.* compared the effects of different normalisation procedures for quantitative LC-MS datasets and concluded that global (or centralised) normalisation and linear regression normalisation worked best in most cases. Notably, the performance of different normalisation procedures varies *per* type of experimental samples and is also different dependent on the criterium that is used to characterise the performance (*i.e.* CV or Local Pooled Error). So preferably, the best procedure should be selected *per* experimental dataset [102]. Careful consideration should be taken whether, and to what threshold, outlier datapoints should be excluded from the calculation of the normalisation factor. Furthermore, scaling of abundance data is important in order to bring the distribution of datapoints more closely to a normal distribution. This is especially important when parametric statistical tests (like Students T-test) are used for difference analysis [103, 104]. In many cases a logarithmic transformation is used for variance stabilisation, see also Fig 5. However, more dedicated log-transformations have been pro-

posed for micro-array data analysis that could also be applied to LC-MS data [105–107]. Anderle *et al*. have tested the variance dependence in relation to peak intensity and developed a two-component (or quadratic) error model. For log-transformed intensities the variance converges to a constant value at high intensities. Whereas at low intensity values the effect of Poisson noise becomes increasingly dominant [61], see also Fig. 5e.

4.7 Data quality evaluation

As mentioned above, matching of (as many as possible) mass features across multiple LC-MS datasets is essential for quantitative comparison. The quality of the used peak detection and alignment algorithms, with respect to aspects as resolution, accuracy, completeness and consistency, is of premium importance for the correctness of the resulting quantitative matching table. So, the quality of these processing steps should be evaluated, in order to provide insight in the correctness of the results and, possibly, to provide clues as to the adjustment of user settings for the different parameters.

Different methods can be used for data evaluation. Visualisation is a powerful tool, be it subjective, for inspection of data quality, where trends, outliers or errors can become evident. More objective clues can be obtained using statistical analysis.



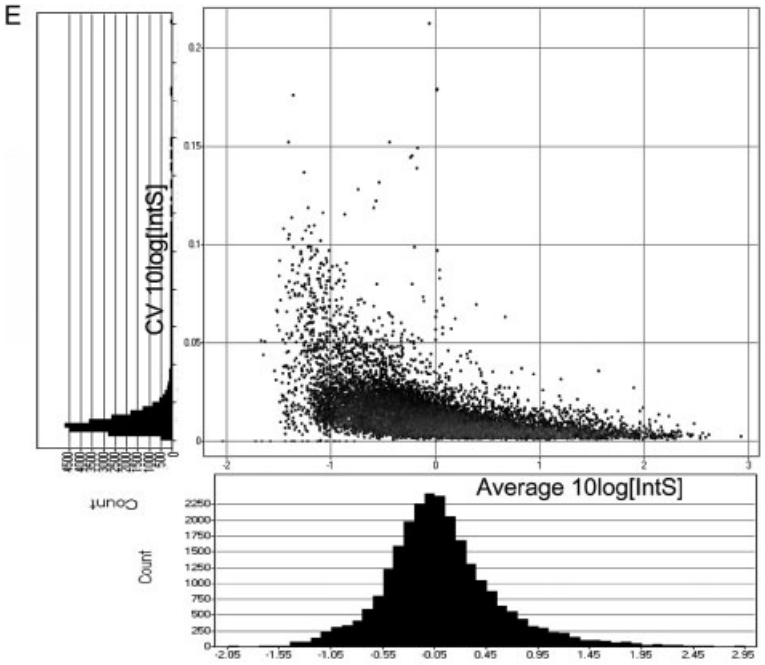


Figure 5. Examples of distribution plots for evaluation of the quantitive reproducibility of replicate LC-MS datasets and evaluation of differences. Peptide mixtures derived from tomato leaf extracellular extracts were analysed in 5-fold replicate on an LTQ-OrbiTrap system and peak frames were detected and quantified with the SIEVE algorithm. The two samples were extracted from two different plants (a resisitant (**R**) *versus* a susceptible (**S**) plant 10 days after inoculation with a plant infectious fungus). A total of 25 160 peaks were detected and quantified by reporting the maximum intensity per peak. NB. Peaks are not de-isotoped nor deconvoluted by SIEVE. Peak intensity data were transformed to ^{10}log scale and centered by subtracting the average ^{10}log value of all peaks per replicate. CV and Students T-test, testing difference between the **R** and **S** sample between five replicates were calculated on the ^{10}log transformed data.
(America *et al.* unpublished data)

(A) Scatterplot of peak intensities of a single **S** LC-MS dataset (run #2) *versus* the average intensities *per* peak frame of the 5-fold replicate **S** dataset. A very narrow distribution along the diagonal is observed, covering *circa* four orders of magnitude in intensity.

(B) Scatterplot of the average peak intensities of 5-fold replicate LC-MS injections of the two different samples **S** *versus* **R**. A much broader distribution is observed, due to the differences between the two samples. A"cloud" of dots can be seen below the diagonal indicating peaks with increased intensity in the susceptible (infected) plant.

(C) Scatterplot of the ratio of peak intensities of samples **S** *versus* **R** plotted against the average replicate intensities of sample **S**. The same cloud of "upregulated peaks" is visible in the top middle of the plot.

(D) Scatterplot of distribution of ^{10}log of the T-test *p*-value *versus* the ratios of peak intensity **S/R**. Clearly, centering of the data (by the chosen method) is not perfect, evidenced by the skewness to the left of the plot. This is also visible in plot **c**.

(E) Distribution plot of the CV and the distribution of average intensity within a 5-fold replicate injection of the sample **S**. The ^{10}log transformation brings the intensity distribution close to gaussian (bottom panel). An inverse relation between peak intensity (x-axis) and CV (y-axis) is observed (middle panel). The peaks with high CV are mostly of very low intensity (left panel).

(note: colour shading from dark to light in all panels is dictated by the ^{10}log[p] of the T-test, as visible in **d**. Figures are produced in Spotfire DecisionSite)

As mentioned in Section 4.3, visualisation of LC-MS peak intensities and detected features provides insight in the quality of actual LC separation as well as the feature detection algorithm, at a global view (Fig. 2). Isotope distribution plots, such as the heat maps provided in MsInspect (Fig. 3), deliver insight in the correctness of charge state detection and help in the selection of filter settings in order to remove low quality assignments.

The quality of alignment can be visualised in an Rt drift plot (Fig.4). Here the shift in Rt between two LC-MS runs is plotted *versus* the Rt (or scan number) for each matched peak pair. Since in most cases, the drift in Rt is equal for all co-eluting peptides, such a plot should display a narrow distribution of data points in a band fluctuating around the zero drift value. A broad distribution indicates non-consistency in peak matching either due to errors in peak matching itself and/or due to errors in peak detection.

Another major quality parameter is the technical reproduciblity of the quantitative approach. Scatterplots of peak intensity data (on a logarithmic scale) within replicate data should display a narrow distribution of datapoints on the diagonal (Fig. 5a). This plot enables easy selection of outlier datapoints in deviating from the (majority of) correctly quantified datapoints that are overlapping on the diagonal line. In

contrast, a scatterplot projecting (within replicates average) peak intensity data between different samples gives a representation of the amount of quantitative differences between these samples (Fig. 5b). A plot of intensity ratio between two samples *versus* the peak intensity indicates whether differential peaks are mainly of low or high intensity (Fig. 5c). This plot also gives a good impression of the quality of the normalisation. The plot should be centered on the middle of the ratio plot. If not, other normalisation procedures should be considered (*e.g.* linear regression). A plot of intensity ratio *versus* the T-test value (here on log scale) also enables a good selection of significant differential peaks, and gives an indication of the normalisation quality (Fig. 5d). This can also be evaluated by analysing the SD of quantitation within replicate injections of identical samples [56, 102]. A plot displaying CV (=SD/average) *versus* the average of peak intensity (within replicates) (Fig. 5e) gives a good impression of the relation between peak intensity and quantitative reproducibility as well as the dynamic range of the quantitative dataset. Not surprisingly, the CV increases at low peak intensities due to a decrease of the S/N. A total distribution histogram of CV values provides a good impression of the total quantitative reproducibility of the complete analysis (Fig. 5e). Outliers in quantitation (or CV) at high peak intensity are mostly due to errors in peak detection or matching.These outliers should be minimised (filtered or preferably corrected, if possible). Outliers at low peak intensity can be due to errors in alignment or in quantitation resulting from low S/N.

Removal or filtering of data points with non-consistent retention time drift values or large CV values from the results list would seem a possibility. However, this may lead to removal of (important) valid datapoints that may negatively affect total data quality (and increase the number of missing data points). Missing values can significantly affect subsequent statistical analysis and machine learning algorithms [78]. Procedures exist to estimate these values as accurately as possible from neighbouring datapoints [108, 109]. A better approach is to optimise peak detection and matching parameters, or even recluster the match table in order to repair errors in peak detection and matching [86]. A manual correction of the match table is not a practical solution, due to the sheer size of the data table (and the risk of subjectivity). Although the use of multiple parameter settings in the alignment software offers the flexibility to adopt and optimise the different processing algorithms to fit with the specific characteristics of the provided LC-MS data, it holds the risk of subjectivity in data analysis. Furthermore, if a large number of parameters need to be set, the effect of each particular parameter on the final data quality may become obscure. This may lead to more or less intuitive trial and error or, even worse, neglecting the parameters (using defaults) which may not be optimal at all. Algorithms where important parameter settings are automatically extracted by data extraction routines (as in PLGS) are preferred. Optimally these settings should then be provided in a parameter table, open for adjustment by the (advanced) user.

4.8 Difference analysis and biomarker detection

If the quality of the match table is satisfactory, the analysis can proceed to perform difference analysis between groups of samples, or classification of samples. Classification of samples is actually based on the collective quantitative behaviour of a large number of peptide abundances, whereas difference analysis uses filtering procedures to select those features that are consistently different in abundance between samples, which are potential biomarkers [52, 104, 110–112]. For difference analysis univariate statistics can be used. Students T-test or ANOVA (on log-transformed abundance data) are often used for selecting differential features [61, 103, 104]. An interesting projection plot is the *p*-value *versus* log(-ratio) plot (see Fig. 5d), as it enables the visual selection of the most significant and most differential features. For univariate statistics analysis, like Students T-test or ANOVA, a correction for multiple hypothesis testing should preferentially be performed. For example, on a list of 10 000 features a *p*-value threshold of 0.01 in theory may result in 100 false positive selected features (in addition to a certain percentage of true positives). To correct for this, statistical methods have been developed to control this False-Dicovery-Rate [18, 113–115].

Advanced machine learning procedures have been reported and compared for classification and difference analysis (feature selection) of MS data sets [111, 116–119]. In many approaches these have been applied to compare MALDI or SELDI data [120–125]. The availability of large scale (many samples) studies with comparative LC-MS is still limited, probably due to the more complicated nature of LC-MS alignment [126]. However, recent studies of multiple replicate LC-MS analyses have assessed the capacity of difference analysis *versus* classification [18, 52, 61]. Listgarten *et al.* demonstrate that classification (*e.g.* Control *versus* spiked, or healthy *versus* diseased) is actually a much easier task than a complete, or perfect, finding of all actual abundance differences. The replicate number of analyses (not unexpectedly) has a strong effect on the precision-recall (true positive rate *versus* false positive rate) ratio [127], with a replicate of five presenting a practical compromise [52]. Clearly, the quality of the alignment procedure in combination with the resolution and precision of the acquired data will have a strong influence on the efficiency of difference detection (precision-recall ratio). So, a performance test should be done in each practical experimental situation.

Depending on the LC-MS method used, it is possible that the identification of the selected differential peptides is already available (as with PMT database, or parallel MS/MS acquisition, see Fig. 1b). In the other scenario where quantitative LC-MS is performed without any knowledge about the identity of the selected features, a targeted MS/MS approach should be followed in order to enable identification of the respective peptides (Fig. 1a). A list of selected features is then made, containing *m/z*, Rt and possibly charge information, which is loaded as "include" or "attention" list into the MS

acquisition control software. Provided that the targeted peptides elute within a user selected window of the given selection, MS/MS acquisition will be triggered to collect fragmentation information from which the sequence of the selected peptide can be derived (by database searching or *de novo* sequencing) [33, 45, 128]. For low abundance peptides, injection of a larger amount of sample may be required in order to provide enough ion intensity for MS/MS detection. Detection sensitivity in MS/MS mode is much less than in MS mode, as the peptide fragments are distributed over a large number of peaks in the MS/MS spectrum. Coupling of peptide MS/MS identification results to previous LC-MS peak abundance data may not be a trivial task, as Rt may drift between the LC-MS runs. Rt normalisation (see Section 4.6) will improve the matching between these different datasets (AMT *vs.* PMT database).

4.9 Multivariate analysis

A clear advantage of the comparative procedure is that it enables the comparison of a large number of LC-MS datasets at a time. Provided that experimental conditions have been optimised for resolution and reproducibility in separation, a large number of samples can be analysed in consecutive runs on the LC-MS system. The LC-MS analysis software will align the resulting datasets as long as there is sufficient overlap in composition of the peptide mixtures, such that enough anchor points for the alignment can be determined. Such multi-sample analysis enables the set-up of time course experiments, or the analysis of a large cohort of experimental groups (*e.g.* patients and control individuals). The quantitative output from such type of analysis can be analysed with multivariate statistics procedures, similar as currently done with micro-array experiments.

Multivariate data analysis has been reported with data from LC-MS experiments in several reports [33, 129–131]. Recently, Prakash *et al.* have presented a software tool, named Chaorder, that implements principles similar to multivariate analysis, in order to assess reproducibility and similarity characteristics of multiple LC-MS data [72]. Based on the alignment algorithm CHAMS [51] a similarity matrix is produced, in which the pairwise similarity of each set of complete raw LC-MS data is represented. A 2-D projection of the (normalised) Euclidian distances between the different LC-MS data provide insight in the similarity metrics between datasets, and as such can visualise (unexpected) biases or trends in data, that may not easily be observed from the complete set of results.

Principal components analysis (PCA) and/or hierarchical clustering provide a good means to visualise the similarities between samples in relation to the differences between classes of samples. PCA provides a multi-dimensional reprojection of the multivariate data, in such a way that the largest level of variation is plotted in the x-dimension. The distance between datapoints is a measure for the variance between the data. Figure 6 displays a PCA plot from

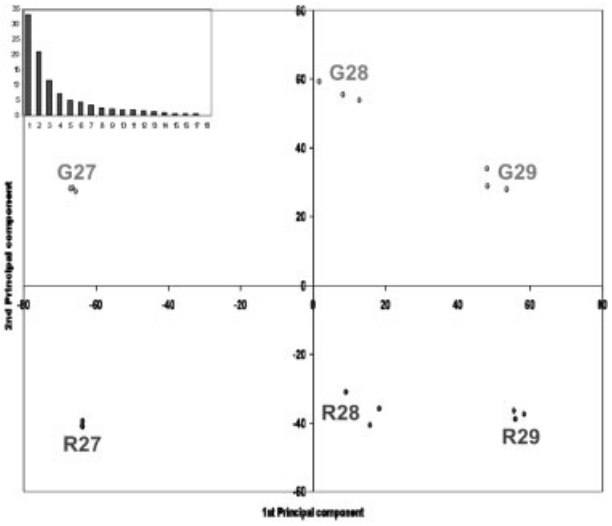


Figure 6. PCA plot based on peptide abundance data collected from 18 aligned RP-LC-MS runs using MetAlign. Three fractions from a SCX fractionation (numbered 27, 28 and 29) were analysed from protein extracts from green (**G**) and red (**R**) tomato. The AMRT data were centered on the average and normalised by dividing by the root mean square value *per* LC-MS dataset. The panel shows the position of each complete LC-MS dataset within the first two principal components. The small histogram insert shows the percentage of total variation covered *per* principal component. A clear separation of the three different SCX fractions and the different plant samples (**G** *versus* **R**) is displayed, while the triplicate datasets cluster very closely together *per* sample (Example taken from [33]).

a SCX fractionation experiment of tryptic peptides from green and red tomato fruits. This example shows that the technical variation between replicate injections is much smaller then the actual variation between the different samples.

Modern variations on PCA exist such as hierarchical PCA [132], or Independent Component Analysis [133–135] which reveal indepent sources of variation within multivariate data. Data from different experiments or different measurements can be integrated via multivariate analysis [136]. Integration of quantitative proteomics data with metabolomics and/or micro-array data provide a promising approach in biomarker discovery and systems biology [137].

Expression plots or heat maps in combination with hierarchical clustering can provide a clustering of multiple peptides that display similar quantitative behaviour within the dataset. It is expected that multiple peptides *per* protein will display very similar quantitative behaviour, and will end in the same clusters [49, 59, 112, 126, 135, 138]. On the other hand, if changes in PTM or multiple isoforms with different expression profiles exist, some peptides from the same protein will end up in a different expression cluster as the majority of its "sister" peptides. As such, comparative LC-MS analysis is not confined to global changes in composition, but also enables the detection of quantitative differences in high detail.

5 Experimental considerations

Comparative LC-MS as quantitative proteomics procedure extracts quantitative information *per* experimental sample out of each indivual LC-MS dataset. This implies that any errors due to experimental and technical deviations will add up to the final quantitative information. This enforces some precautions in experimental design in order to guarantee optimal quality of the results from which significant conclusions can be drawn. General considerations for the design of experiment should be considered, just as for other (quantitative) experimental procedures [106, 115, 123]. Any experimental sample preparation procedure should be performed with quantitative reproducibility, as much as possible. The use of internal standard proteins, which can be spiked at an early point of sample preparation, will provide reference points in the final dataset, on which the matching as well as quantitative information can be calibrated. Furthermore, the acquisition of quantitative LC-MS data should preferably be performed in a consecutive series of measurements, in order to minimise drifts in (chromatographic) separation as well as (MS) detection. Control LC-MS runs (*i.e.*. with only the spike sample) should be performed intermittently between experimental runs in order to enable quality control during the complete series of analysis [72]. Comparison of datasets that have been acquired at different time points (*e.g.* separated by weeks) can be quite limited due to unexpected changes in separation conditions (like temperature, buffer composition or even column condition). The use of internal reference markers and also the use of fragmentation information as anchor points for Rt normalisation will provide more tolerance and accuracy in matching. While comparative LC-MS, often referred to as label-free LC-MS, enables quantitative analysis without the use of stable-isotope labeling, it actually can profit from the use of a stable isotope labeled sample as an internal reference. Such an internal reference could be a standardised control sample, mixed with each experimental sample. Its use will provide a large number of reference points to which mass, Rt and peak intensity can be calibrated throughout the LC-MS datasets. While using the high resolution feature detection algorithms developed for comparative LC-MS the quantitative accuracy will improve even more, with this high density reference points.

How complex can a protein sample be to be analysed by comparative LC-MS? As mentioned before, the resolution in separation and MS detection both have a strong influence on the total peak capacity. So, ultra-high resolution will have the strongest detection power [12, 45, 139]. It is difficult to present a general advice. In our experiment, displayed in Fig. 5, *circa* 25 000 peaks (non-deisotoped nor deconvoluted) were detected from a relatively "simple" extracellular extract, revealing *circa* 300 spots on a 2-D gel (data not shown). When deisotoped and deconvoluted, these data would approximately point to 5000 peptide features. More complex protein mixtures can produce more then 100 000 peptides. These cannot all be resolved in a single 1-D LC-MS run. A compromise should be decided upon between the enhanced peak detection of multidimensional LC separation upfront of the MS analysis [138], *versus* the associated increase in number of LC-MS runs (with the associated increase in data complexity). On the other hand, succesfull comparative biomarker studies have been reported from complete human serum samples, without further enrichment, using nanoLC-QTOF equipment [59]. Here the more intense peptide features will dominate the analysis, which does not necessarily preclude success. As in many -omics experiments, difference detection and classification does not imply that we detect all differences. This is acceptable as long as the detected differences are significant and reproducible.

6 Conclusion

Comparative LC-MS is a relatively recent approach in quantitative proteomics analysis. Software tools, which are essential for this procedure, are still in development and have room for improvement. User interface and calculation efficiency of the processing workflow can be improved in many cases. More importantly, the integration of data at the level of MS and MS/MS from multiple experiments needs more transparent implementation in database systems. The tolerances for drift, especially in the Rt domain, are (in most cases) tight, which enforce a relatively strong control of separation conditions in the experimental set-up. On the other hand, the high accuracy in quantitative determination of peptide abundances at an impressive level of detail requires an experimental set-up where quantitative aspects like protein extraction efficiencies and sample stability need to be well controlled. At the current state already, it appears that the quantitative reproducibility of technical replicates is in many cases much better than experimental reproducibility [61, 140].

In conclusion, with improvements in the efficiency of the data analysis workflow, comparative LC-MS has a tremendous potential for application in biomarker detection analysis due to its high level of detail, its quantitative accuracy and its capacity for large scale applications. Successful applications in biomedical biomarker discovery have recently been reported in the field of breast cancer [126, 138], Gaucher disease [59] and proteinurea [131]. We are truly confident that many will follow.

We acknowledge the collaboration with Dr. Hans Vissers of Waters Corporation for the UPLC-QTOF data and evaluation of the Expression Informatics module of ProteinLynx Global Server, and the collaboration with Dr. Wilfried Voorhorst of Thermo Fisher Scientific for the LTQ-OrbiTrap data and evaluation of the SIEVE software.

We acknowledge the funding of the projects "Quantitative proteomics" for Antoine America and Jan H. G. Cordewener as part of the Netherlands Proteomics Centre and the Centre for

Biosystems and Genomics, both of which are part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

The authors have declared no conflict of interest.

7 References

[1] Veenstra, T. D., Conrads, T. P., Hood, B. L., Avellino, A. M. *et al.*, Biomarkers: Mining the biofluid proteome. *Mol. Cell. Proteomics* 2005, *4*, 409–418.

[2] Ackermann, B. L., Hale, J. E., Duffin, K. L., The role of mass spectrometry in biomarker discovery and measurement. *Curr. Drug Metab.* 2006, *7*, 525–539.

[3] Kussmann, M., Raymond, F., Affolter, M., OMICS-driven biomarker discovery in nutrition and health. *J. Biotechnol.* 2006, *124*, 758–787.

[4] Powell, D. W., Merchant, M. L., Link, A. J., Discovery of regulatory molecular events and biomarkers using 2D capillary chromatography and mass spectrometry. *Expert Rev. Proteomics* 2006, *3*, 63–74.

[5] Whiteaker, J. R., Zhang, H., Eng, J. K., Fang, R. *et al.*, Head-to-head comparison of serum fractionation techniques. *J. Proteome Res.* 2007, *6*, 828–836.

[6] Stasyk, T., Huber, L. A., Zooming in: fractionation strategies in proteomics. *Proteomics* 2004, *4*, 3704–3716.

[7] Sharma, S., Simpson, D. C., Tolic, N., Jaitly, N. *et al.*, Proteomic profiling of intact proteins using WAX-RPLC 2-D separations and FTICR mass spectrometry. *J. Proteome Res.* 2007, *6*, 602–610.

[8] Sheng, S., Chen, D., Van Eyk, J. E., Multidimensional liquid chromatography separation of intact proteins by chromatographic focusing and reversed phase of the human serum proteome: optimization and protein database. *Mol. Cell. Proteomics* 2006, *5*, 26–34.

[9] Chen, E. I., Hewel, J., Felding-Habermann, B., Yates, J. R., 3rd, Large scale protein profiling by combination of protein fractionation and multidimensional protein identification technology (MudPIT). *Mol. Cell. Proteomics* 2006, *5*, 53–56.

[10] Wolters, D. A., Washburn, M. P., Yates, J. R., 3rd, An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* 2001, *73*, 5683–5690.

[11] Washburn, M. P., Wolters, D., Yates, J. R., 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* 2001, *19*, 242–247.

[12] Shen, Y., Jacobs, J. M., Camp, D. G., 2nd,Fang, R. *et al.*, Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. *Anal. Chem.* 2004, *76*, 1134–1144.

[13] Roy, S. M., Anderle, M., Lin, H., Becker, C. H., Differential expression profiling of serum proteins and metabolites for biomarker discovery. *Int. J. Mass Spectrom.* 2004, *238*, 163–171.

[14] Higgs, R. E., Knierman, M. D., Gelfanova, V., Butler, J. P., Hale, J. E., Comprehensive label-free method for the relative quantification of proteins from biological samples. *J. Proteome Res.* 2005, *4*, 1442–1450.

[15] Silva, J. C., Denny, R., Dorschel, C. A., Gorenstein, M. *et al.*, Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* 2005, *77*, 2187–2200.

[16] Silva, J. C., Denny, R., Dorschel, C., Gorenstein, M. V. *et al.*, Simultaneous qualitative and quantitative analysis of the Escherichia coli proteome: A sweet tale. *Mol. Cell. Proteomics* 2006, *5*, 589–607.

[17] Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G. *et al.*, Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* 2005, *4*, 1487–1502.

[18] Cho, H., Smalley, D. M., Theodorescu, D., Ley, K., Lee, J. K., Statistical identification of differentially labeled peptides from liquid chromatography tandem mass spectrometry. *Proteomics* 2007, *7*, 3681–3692.

[19] Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B. *et al.*, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* 2002, *1*, 376–386.

[20] Flory, M. R., Griffin, T. J., Martin, D., Aebersold, R., Advances in quantitative proteomics using stable isotope tags. *Trends Biotechnol.* 2002, *20*, S23–S29.

[21] Moritz, B., Meyer, H. E., Approaches for the quantification of protein concentration ratios. *Proteomics* 2003, *3*, 2208–2220.

[22] Julka, S., Regnier Fred, E., Recent advancements in differential proteomics based on stable isotope coding. *Brief. Funct. Genomic. Proteomic.* 2005, *4*, 158–177.

[23] Moulder, R., Filen, J. J., Salmi, J., Katajamaa, M. *et al.*, A comparative evaluation of software for the analysis of liquid chromatography-tandem mass spectrometry data from isotope coded affinity tag experiments. *Proteomics* 2005, *5*, 2748–2760.

[24] Li, X. J., Zhang, H., Ranish, J. A., Aebersold, R., Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal. Chem.* 2003, *75*, 6648–6657.

[25] Zhang, X., Hines, W., Adamec, J., Asara, J. M. *et al.*, An automated method for the analysis of stable isotope labeling data in proteomics. *J. Am. Soc. Mass Spectrom.* 2005, *16*, 1181–1191.

[26] Krijgsveld, J., Ketting, R. F., Mahmoudi, T., Johansen, J. *et al.*, Metabolic labeling of *C. elegans* and *D. melanogaster* for quantitative proteomics. *Nat. Biotechnol.* 2003, *21*, 927–931.

[27] Gehrmann, M. L., Hathout, Y., Fenselau, C., Evaluation of metabolic labeling for comparative proteomics in breast cancer cells. *J. Proteome Res.* 2004, *3*, 1063–1068.

[28] Wu, C. C., MacCoss, M. J., Howell, K. E., Matthews, D. E., Yates, J. R., 3rd, Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. *Anal. Chem.* 2004, *76*, 4951–4959.

[29] Beynon, R. J., Pratt, J. M., Metabolic labeling of proteins for proteomics. *Mol. Cell. Proteomics* 2005, *4*, 857–872.

[30] Lafaye, A., Labarre, J., Tabet, J. C., Ezan, E., Junot, C., Liquid chromatography-mass spectrometry and 15N metabolic labeling for quantitative metabolic profiling. *Anal. Chem.* 2005, *77*, 2026–2033.

[31] Snijders, A. P., de Vos, M. G., Wright, P. C., Novel approach for peptide quantitation and sequencing based on 15N and 13C metabolic labeling. *J. Proteome Res.* 2005, *4*, 578–585.

[32] Tong, W., Link, A., Eng, J. K., Yates, J. R., 3rd, Identification of proteins in complexes by solid-phase microextraction/multistep elution/capillary electrophoresis/tandem mass spectrometry. *Anal. Chem.* 1999, *71*, 2270–2278.

[33] America, A. H. P., Cordewener, J. H. G., van Geffen, M. H. A., Lommen, A. *et al.*, Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional LC-MS. *Proteomics* 2006, *6*, 641–653.

[34] Liu, H., Sadygov, R. G., Yates, J. R., A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 2004, *76*, 4193–4201.

[35] Ishihama, Y., Oda, Y., Tabata, T., Sato, T. *et al.*, Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides *per* protein. *Mol. Cell. Proteomics* 2005, *4*, 1265–1272.

[36] Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K. *et al.*, Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J. Proteome Res.* 2006, *5*, 2339–2347.

[37] Paoletti, A. C., Parmely, T. J., Tomomori-Sato, C., Sato, S. *et al.*, Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc. Natl. Acad. Sci. USA* 2006, *103*, 18928–18933.

[38] Wiener, M. C., Sachs, J. R., Deyanova, E. G., Yates, N. A., Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. *Anal. Chem.* 2004, *76*, 6085–6096.

[39] Radulovic, D., Jelveh, S., Ryu, S., Hamilton, T. G. *et al.*, Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* 2004, *3*, 984–997.

[40] Pasa-Tolic, L., Masselon, C., Barry, R. C., Shen, Y., Smith, R. D., Proteomic analyses using an accurate mass and time tag strategy. *Biotechniques* 2004, *37*, 621–624, 626–633, 636.

[41] Bellew, M., Coram, M., Fitzgibbon, M., Igra, M. *et al.*, A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* 2006, *22*, 1902–1909.

[42] Katajamaa, M., Oresic, M., Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics* 2005, *6*, 179.

[43] Katajamaa, M., Miettinen, J., Oresic, M., MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 2006, *22*, 634–636.

[44] Katajamaa, M., Oresic, M., Data processing for mass spectrometry-based metabolomics. *J. Chromatogr. A* 2007, *1158*, 318–328.

[45] Fang, R., Elias, D. A., Monroe, M. E., Shen, Y. *et al.*, Differential label-free quantitative proteomic analysis of *Shewanella oneidensis* cultured under aerobic and suboxic conditions by accurate mass and time tag approach. *Mol. Cell. Proteomics* 2006, *5*, 714–725.

[46] Fischer, B., Grossmann, J., Roth, V., Gruissem, W. *et al.*, Semi-supervised LC/MS alignment for differential proteomics. *Bioinformatics* 2006, *22*, e132–140.

[47] Ono, M., Shitashige, M., Honda, K., Isobe, T. *et al.*, Label-free quantitative proteomics using large peptide data sets generated by nanoflow liquid chromatography and mass spectrometry. *Mol. Cell. Proteomics* 2006, *5*, 1338–1347.

[48] Wang, G., Wu, W. W., Zeng, W., Chou, C. L., Shen, R. F., Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: Reproducibility, linearity, and application with complex proteomes. *J. Proteome Res.* 2006, *5*, 1214–1223.

[49] Li, X. J., Yi, E. C., Kemp, C. J., Zhang, H., Aebersold, R., A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol. Cell. Proteomics* 2005, *4*, 1328–1340.

[50] Johansson, C., Samskog, J., Sundstrom, L., Wadensten, H. *et al.*, Differential expression analysis of Escherichia coli proteins using a novel software for relative quantitation of LC-MS/MS data. *Proteomics* 2006, *6*, 4475–4485.

[51] Prakash, A., Mallick, P., Whiteaker, J., Zhang, H. *et al.*, Signal maps for mass spectrometry-based comparative proteomics. *Mol. Cell. Proteomics* 2006, *5*, 423–432.

[52] Listgarten, J., Neal, R. M., Roweis, S. T., Wong, P., Emili, A., Difference detection in LC-MS data for protein biomarker discovery. *Bioinformatics* 2007, *23*, e198–204.

[53] Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S. *et al.*, SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* 2007, *7*, 3470–3480.

[54] Schmidt, A., Karas, M., Dulcks, T., Effect of different solution flow rates on analyte ion signals in nano-ESI MS, or: when does ESI turn into nano-ESI? *J. Am. Soc. Mass Spectrom.* 2003, *14*, 492–500.

[55] Shen, Y., Zhao, R., Berger, S. J., Anderson, G. A. *et al.*, High-efficiency nanoscale liquid chromatography coupled on-line with mass spectrometry using nanoelectrospray ionization for proteomics. *Anal. Chem.* 2002, *74*, 4235–4249.

[56] Listgarten, J., Emili, A., Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* 2005, *4*, 419–434.

[57] Zimmer, J. S., Monroe, M. E., Qian, W. J., Smith, R. D., Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom. Rev.* 2006, *25*, 450–482.

[58] Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P., Geromanos, S. J., Absolute quantification of proteins by LCMSE: A virtue of parallel MS acquisition. *Mol. Cell. Proteomics* 2006, *5*, 144–156.

[59] Vissers, J. P., Langridge, J. I., Aerts, J. M., Analysis and quantification of diagnostic serum markers and protein signatures for Gaucher disease. *Mol. Cell. Proteomics* 2007, *6*, 755–766.

[60] Jaffe, J. D., Mani, D. R., Leptos, K. C., Church, G. M. *et al.*, PEPPeR, a platform for experimental proteomic pattern recognition. *Mol. Cell. Proteomics* 2006, *5*, 1927–1941.

[61] Anderle, M., Roy, S., Lin, H., Becker, C., Joho, K., Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics* 2004, *20*, 3575–3582.

[62] Hastings, C. A., Norton, S. M., Roy, S., New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data. *Rapid Commun. Mass Spectrom.* 2002, *16*, 462–467.

[63] Roy, S. M., Becker, C. H., Quantification of proteins and metabolites by mass spectrometry without isotopic labeling. *Methods Mol. Biol.* 2007, *359*, 87–105.

[64] Palagi, P. M., Walther, D., Quadroni, M., Catherinet, S. *et al.*, MSight: An image analysis software for liquid chromatography-mass spectrometry. *Proteomics* 2005, *5*, 2381–2384.

[65] May, D., Fitzgibbon, M., Liu, Y., Holzman, T. *et al.*, A platform for accurate mass and time analyses of mass spectrometry data. *J. Proteome Res.* 2007, *6*, 2685–2694.

[66] Leptos, K. C., Sarracino, D. A., Jaffe, J. D., Krastins, B., Church, G. M., MapQuant: Open-source software for large-scale protein quantification. *Proteomics* 2006, *6*, 1770–1782.

[67] Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L. *et al.*, An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* 2002, *2*, 513–523.

[68] Kohlbacher, O., Reinert, K., Gropl, C., Lange, E. *et al.*, TOPP–the OpenMS proteomics pipeline. *Bioinformatics* 2007, *23*, e191–197.

[69] Lange, E., Gropl, C., Reinert, K., Kohlbacher, O., Hildebrandt, A., High-accuracy peak picking of proteomics data using wavelet techniques. *Pac. Symp. Biocomput.* 2006, 243–254.

[70] Lange, E., Gropl, C., Schulz-Trieglaff, O., Leinenbach, A. *et al.*, A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics* 2007, *23*, i273–281.

[71] Zhang, X., Asara, J. M., Adamec, J., Ouzzani, M., Elmagarmid, A. K., Data pre-processing in liquid chromatography-mass spectrometry-based proteomics. *Bioinformatics* 2005, *21*, 4054–4059.

[72] Prakash, A., Piening, B., Whiteaker, J., Zhang, H. *et al.*, Assessing bias in experiment design for large scale mass spectrometry-based Quantitative proteomics. *Mol. Cell. Proteomics* 2007, *6*, 1741–1748.

[73] Prince, J. T., Marcotte, E. M., Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Anal. Chem.* 2006, *78*, 6140–6152.

[74] Jaitly, N., Monroe, M. E., Petyuk, V. A., Clauss, T. R. *et al.*, Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal. Chem.* 2006, *78*, 7397–7409.

[75] Du, P., Angeletti, R. H., Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution. *Anal. Chem.* 2006, *78*, 3385–3392.

[76] Du, P., Sudha, R., Prystowsky, M. B., Angeletti, R. H., Data reduction of isotope-resolved LC-MS spectra. *Bioinformatics* 2007, *23*, 1394–1400.

[77] Wang, P., Tang, H., Fitzgibbon, M. P.,McIntosh, M. *et al.*, A statistical method for chromatographic alignment of LC-MS data. *Biostatistics* 2007, *8*, 357–367.

[78] Wang, P., Tang, H., Zhang, H., Whiteaker, J. *et al.*, Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pac. Symp. Biocomput.* 2006, *11*, 315–326.

[79] Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M. *et al.*, A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* 2004, *22*, 1459–1466.

[80] mzXML, http://tools.proteomecenter.org/mzXMLschema.php

[81] mzML, http://www.psidev.info/index.php?q=node/257

[82] Du, P., Kibbe, W. A., Lin, S. M., Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* 2006, *22*, 2059–2065.

[83] Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A., Coombes, K. R., Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics* 2007, published on-line 20 Sep 2007, DOI 10.1111/j.1541-042.2007.00895.

[84] Windig, W., Smith, W. F., Chemometric analysis of complex hyphenated data. Improvements of the component detection algorithm. *J. Chromatogr. A* 2007, *1158*, 251–257.

[85] Zhang, Z., Marshall, A. G., A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J. Am. Soc. Mass Spectrom.* 1998, *9*, 225–233.

[86] Groot, J. C. W. d., Fiers, M. W. E. J., Ham, R. C. H. J. v., America, A. H. P., Post-aligment Clustering Procedure (PACP) for comparative quantitative proteomics LC-MS data. *Proteomics* 2008, *8*, 32–36.

[87] Piening, B. D., Wang, P., Bangur, C. S., Whiteaker, J. *et al.*, Quality control metrics for LC-MS feature detection tools demonstrated on *Saccharomyces cerevisiae* proteomic profiles. *J. Proteome Res.* 2006, *5*, 1527–1534.

[88] Bylund, D., Danielsson, R., Malmquist, G., Markides, K. E., Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *J. Chromatogr. A* 2002, *961*, 237–244.

[89] Lipton, M. S., Pasa-Tolic, L., Anderson, G. A., Anderson, D. J. *et al.*, Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. *Proc. Natl. Acad. Sci. USA* 2002, *99*, 11049–11054.

[90] R-project, www.r-project.org

[91] DecisionSite, http://www.spotfire.com/products/decisionsite.cfm

[92] GeneMath, http://www.applied-maths.com/genemaths/genemaths.htm

[93] SAM, http://www-stat.stanford.edu/~tibs/SAM/

[94] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S. *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 2005, *102*, 15545–15550.

[95] Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W. *et al.*, TM4 microarray software suite. *Methods Enzymol.* 2006, *411*, 134–193.

[96] Sturn, A., Quackenbush, J., Trajanoski, Z., Genesis: Cluster analysis of microarray data. *Bioinformatics* 2002, *18*, 207–208.

[97] Storey, J. D., Tibshirani, R., Statistical methods for identifying differentially expressed genes in DNA microarrays. *Methods Mol. Biol.* 2003, *224*, 149–157.

[98] Strittmatter, E. F., Kangas, L. J., Petritis, K., Mottaz, H. M. *et al.*, Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *J. Proteome Res.* 2004, *3*, 760–769.

[99] Krokhin, O. V., Craig, R., Spicer, V., Ens, W. *et al.*, An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: Its application to

protein peptide mapping by off-line HPLC-MALDI MS. *Mol. Cell. Proteomics* 2004, *3*, 908–919.

[100] Petritis, K., Kangas, L. J., Yan, B., Monroe, M. E. *et al.*, Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal. Chem.* 2006, *78*, 5026–5039.

[101] Klammer, A. A., Yi, X., MacCoss, M. J., Noble, W. S., Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. *Anal. Chem.* 2007, *79*, 6111–6118.

[102] Callister, S. J., Barry, R. C., Adkins, J. N., Johnson, E. T. *et al.*, Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.* 2006, *5*, 277–286.

[103] Karp, N. A., Lilley, K. S., Design and analysis issues in quantitative proteomics studies. *Proteomics* 2007, *7*, 42–50.

[104] Urfer, W., Grzegorczyk, M., Jung, K., Statistics for proteomics: A review of tools for analyzing experimental data. *Proteomics* 2006, *6*, 48–55.

[105] Durbin, B. P., Rocke, D. M., Variance-stabilizing transformations for two-color microarrays. *Bioinformatics* 2004, *20*, 660–667.

[106] Rocke, D. M., Design and analysis of experiments with high throughput biological assay data. *Semin. Cell Dev. Biol.* 2004, *15*, 703–713.

[107] Quackenbush, J., Microarray data normalization and transformation. *Nat. Genet.* 2002, *32*, 496–501.

[108] Sehgal, M. S., Gondal, I., Dooley, L. S., Collateral missing value imputation: A new robust missing value estimation algorithm for microarray data. *Bioinformatics* 2005, *21*, 2417–2423.

[109] Kim, H., Golub, G. H., Park, H., Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* 2005, *21*, 187–198.

[110] Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* 2002, *99*, 6567–6572.

[111] Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S. *et al.*, Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics* 2004, *20*, 3034–3044.

[112] Varshavsky, R., Gottlieb, A., Linial, M., Horn, D., Novel unsupervised feature filtering of biological data. *Bioinformatics* 2006, *22*, e507–513.

[113] Dudoit, S., Shaffer, J. P., Boldrick, J. C., Multiple hypothesis Ttesting in microarray experiments. *Stat. Sci.* 2003, *18*, 71–103.

[114] Storey, J. D., Tibshirani, R., Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* 2003, *100*, 9440–9445.

[115] Karp, N. A., McCormick, P. S., Russell, M. R., Lilley, K. S., Experimental and statistical considerations to avoid false conclusions in proteomics studies using differential in-gel electrophoresis. *Mol. Cell. Proteomics* 2007, *6*, 1354–1364.

[116] Geurts, P., Fillet, M., de Seny, D., Meuwis, M. A. *et al.*, Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics* 2005, *21*, 3138–3145.

[117] Levner, I., Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics* 2005, *6*, 68.

[118] Zhang, X., Lu, X., Shi, Q., Xu, X. Q. *et al.*, Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics* 2006, *7*, 197.

[119] Hendriks, M. M., Smit, S., Akkermans, W. L., Reijmers, T. H. *et al.*, How to distinguish healthy from diseased? Classification strategy for mass spectrometry-based clinical proteomics. *Proteomics* 2007, *7*, 3672–3680.

[120] Belluco, C., Petricoin, E. F., Mammano, E., Facchiano, F. *et al.*, Serum proteomic analysis identifies a highly sensitive and specific discriminatory pattern in stage 1 breast cancer. *Ann. Surg. Oncol.* 2007, *14*, 2470–2476.

[121] Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J. *et al.*, Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002, *359*, 572–577.

[122] Petricoin, E. F., 3rd, Ornstein, D. K., Paweletz, C. P., Ardekani, A. *et al.*, Serum proteomic patterns for detection of prostate cancer. *J. Natl. Cancer Inst.* 2002, *94*, 1576–1578.

[123] Hu, J., Coombes, K. R., Morris, J. S., Baggerly, K. A., The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief. Funct. Genomic. Proteomic.* 2005, *3*, 322–331.

[124] Yasui, Y., Pepe, M., Hsu, L., Adam, B. L., Feng, Z., Partially supervised learning using an EM-boosting algorithm. *Biometrics* 2004, *60*, 199–206.

[125] Wagner, M., Naik, D., Pothen, A., Protocols for disease classification from mass spectrometry data. *Proteomics* 2003, *3*, 1692–1698.

[126] Ru, Q. C., Zhu, L. A., Silberman, J., Shriver, C. D., Label-free semiquantitative peptide feature profiling of human breast cancer and breast disease sera via two-dimensional liquid chromatography-mass spectrometry. *Mol. Cell. Proteomics* 2006, *5*, 1095–1104.

[127] Karp, N. A., Spencer, M., Lindsay, H., O'Dell, K., Lilley, K. S., Impact of replicate types on proteomic expression analysis. *J. Proteome Res.* 2005, *4*, 1867–1871.

[128] Masselon, C., Pasa-Tolic, L., Tolic, N., Anderson, G. A. *et al.*, Targeted comparative proteomics by liquid chromatography-tandem Fourier ion cyclotron resonance mass spectrometry. *Anal. Chem.* 2005, *77*, 400–406.

[129] Weckwerth, W., Wenzel, K., Fiehn, O., Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics* 2004, *4*, 78–83.

[130] Gaspari, M., Verhoeckx, K. C., Verheij, E. R., van der Greef, J., Integration of two-dimensional LC-MS with multivariate statistics for comparative analysis of proteomic samples. *Anal. Chem.* 2006, *78*, 2286–2296.

[131] Kemperman, R. F., Horvatovich, P. L., Hoekman, B., Reijmers, T. H. *et al.*, Comparative urine analysis by liquid chromatography-mass spectrometry and multivariate statistics: Method development, evaluation, and application to proteinuria. *J. Proteome Res.* 2007, *6*, 194–206.

[132] Kong, S. W., Pu, W. T., Park, P. J., A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics* 2006, *22*, 2373–2380.

[133] Scholz, M., Gatzek, S., Sterling, A., Fiehn, O., Selbig, J., Metabolite fingerprinting: Detecting biological features by independent component analysis. *Bioinformatics* 2004, *20*, 2447–2454.

[134] Scholz, M., Selbig, J., Visualization and analysis of molecular data. *Methods Mol. Biol.* 2007, *358*, 87–104.

[135] Steuer, R., Morgenthal, K., Weckwerth, W., Selbig, J., A gentle guide to the analysis of metabolomic data. *Methods Mol. Biol.* 2007, *358*, 105–126.

[136] Weckwerth, W., Metabolomics: from pattern recognition to biological interpretation. *Drug Discov. Today* 2005, *10*, 1551–1558.

[137] Searls, D. B., Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* 2005, *4*, 45–58.

[138] Patwardhan, A. J., Strittmatter, E. F., Camp, D. G., 2nd, Smith, R. D., Pallavicini, M. G., Quantitative proteome analysis of breast cancer cell lines using 18O-labeling and an accurate mass and time tag strategy. *Proteomics* 2006, *6*, 2903–2915.

[139] Shen, Y., Tolic, N., Masselon, C., Pasa-Tolic, L. *et al.*, Ultrasensitive proteomics using high-efficiency on-line micro-SPE-nanoLC-nanoESI MS and MS/MS. *Anal. Chem.* 2004, *76*, 144–154.

[140] Wang, W., Zhou, H., Lin, H., Roy, S. *et al.*, Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* 2003, *75*, 4818–4826.

