Amino Acids (2011) 40:697–711
DOI 10.1007/s00726-010-0704-2

ORIGINAL ARTICLE

Observed peptide p*I* and retention time shifts as a result of post-translational modifications in multidimensional separations using narrow-range IPG-IEF

Johan Lengqvist · Hanna Eriksson · Marcus Gry · Kristina Uhlén · Christina Björklund · Bengt Bjellqvist · Per-Johan Jakobsson · Janne Lehtiö

Received: 26 February 2010 / Accepted: 14 July 2010 / Published online: 20 August 2010

Abstract Modified peptides constitute a sub-population among the tryptic peptides analyzed in LC–MS based shotgun proteomics experiments. For larger proteomes including the human proteome, the tryptic peptide pool is very large, which necessitates some form of sample fractionation. By carefully choosing the sample fractionation and separation methods applied as shown here for the combination of narrow-range immobilized pH gradient isoelectric focusing (IPG-IEF) and nanoUPLC–MS, significantly increased information content can be achieved. Relatively low standard deviations were obtained for such multidimensional separations in terms of peptide p*I* (<0.05 p*I* units) and retention time (<0.3 min for a 350 min gradient) for a selection of highly complex proteomics samples. Using narrow-range IPG-IEF, experimental and predicted p*I* were in relative good agreement. However, based on our data, retention time prediction algorithms need further improvements in accuracy to match state-of-the-art reversed-phase chromatography performance. General trends of peptide p*I* shifts induced by common modifications including deamidations and N-terminal modifications are described. Deamidations of glutamine and asparagines shift peptide p*I* by approximately 1.5 p*I* units, making the peptides more acidic. Additionally, a novel p*I* shift (+ ~ 0.4 p*I* units) was found associated with dethiomethyl Met modifications. Further, the effects of these modifications as well as methionine oxidation were investigated in terms of experimentally observed retention time shifts in the chromatographic separation step. Clearly, post-translational modification-induced influences on peptide p*I* and retention time can be accurately and reproducibly measured using narrow-range IPG-IEF and high-performance nanoLC–MS. Even at modest mass accuracy (±50 ppm), the inclusion of peptide p*I* (±0.2 p*I* units) and/or retention time (±20 min) criteria are highly informative for human proteome analyses. The applications of using this information to identify post-translationally modified peptides and improve data analysis workflows are discussed.

Keywords Isoelectric focusing · Post-translational modifications · Retention time

Electronic supplementary material The online version of this article (doi:10.1007/s00726-010-0704-2) contains supplementary material, which is available to authorized users.

J. Lengqvist · M. Gry · C. Björklund
Department of Safety Assessment, Molecular Toxicology,
AstraZeneca R&D, 15185 Södertälje, Sweden

J. Lengqvist · H. Eriksson · P.-J. Jakobsson · J. Lehtiö
Karolinska Biomics Center, Karolinska University Hospital,
17176 Stockholm, Sweden

J. Lehtiö (✉)
Department of Oncology and Pathology, Karolinska Biomics
Center (Z5:02), Karolinska Institutet, 17176 Stockholm, Sweden
e-mail: janne.lehtio@ki.se

H. Eriksson · P.-J. Jakobsson
Rheumatology Unit, Department of Medicine, Karolinska
Institutet, 17176 Stockholm, Sweden

K. Uhlén · B. Bjellqvist
GE Healthcare Bio-Sciences AB, 75184 Uppsala, Sweden

Introduction

Shotgun proteomics has to deal with extreme sample complexity after the entire cellular or plasma protein content has been digested into peptides. The 20,000–25,000 human gene products will yield ~2,000,000 tryptic peptides if one missed cleavage site is allowed (2,045,133 for the 21,273 human proteins of the SwissProt database). Additionally, this number counts only non-modified peptides. Presently, about 500 amino acid modifications are known and together with alternative splice variants and amino acid substitutions they make up the true complexity of the proteome (Ahrne et al. 2009). Protein identification in these proteomic experiments is based on peptide fragmentation data through multiple MS/MS runs. Using database search software, assignment of a likely peptide sequence to each fragment ion spectrum is attempted and a match score is given to indicate the quality (i.e., the confidence) of the assignment. For a typical LC–MS/MS experiment, 5–30% of the acquired MS/MS spectra will be assigned to a peptide sequence (Ahrne et al. 2009; Nielsen et al. 2006). To increase the success rate, additional runs of the same sample may be required. One of the major problems is the complexity derived from post-translationally modified (PTM) peptides. For a typical search engine software such as Mascot, the peptide modifications to consider in a search have to be determined a priori. Allowing several modifications increases the search space leading to longer search times, effectively limiting the number of modifications that can be considered [for a discussion of the problem, see e.g., (Fenyö and Beavis 2008)]. As a result, modified peptides, although obviously present, form a much less well-characterized sub-population in shotgun proteomics.

Non-biased (global) PTM detection using so-called open-modification search engines is starting to appear to cope with the problem of inadequate peptide PTM data generation (Tanner et al. 2005; Matthiesen et al. 2005; Gupta et al. 2007; Savitski et al. 2006). The major constraint with open-modification searches is that each spectrum takes longer to process, which necessitates an initial database reduction step (Ahrne et al. 2009). As excellently reviewed by Ahrne et al. (Ahrne et al. 2009), database reduction can be done through so-called sequence tag-based approaches or through multiple rounds of searching. Apart from being relatively slow and insensitive, an open-modification search is also more error prone than a traditional one. Thus, some form of hit list curation is needed. One option is manual inspection of each MS/MS spectrum, but this requires time and experience making it virtually impossible for large MS/MS data sets. Experimentally derived information such as MS/MS/MS (MS^3) sequencing is another option to support or disqualify individual peptide assignments (Olsen and Mann 2004). Further, the measured precursor ion mass, especially when measured with high mass accuracy, is very informative and it is the primary database search constraint to limit the search space (Zubarev and Mann 2007). Another readily available experimentally obtained piece of information that can also be used is the reversed-phase liquid chromatography (RPLC) retention time. Additionally, if the peptide mixture has been separated using some form of isoelectric focusing (IEF), the peptide p*I* value can be used (Essader et al. 2005; Krijgsveld et al. 2006). The accuracy of modern two-dimensional IEF and RPLC technology prompted us to investigate the information content of experimentally derived mass, p*I* and retention time information, and the impact of peptide modifications on these separations.

Apart from being an excellent means of separating a complex mixture of peptides, RPLC can provide additional information that can aid in the identification process as outlined above. The chromatographic retention time can be considered as a structure-dependent parameter that is constant for a peptide at specific separation conditions. Different prediction algorithms have been developed, the early ones relying on the hydrophobicity of the peptide as a sum of the hydrophobicities of all the constituent amino acids (Meek 1980; Browne et al. 1982). However, it was soon realized that the existing algorithms needed improvements; more specifically, they should consider other peptide properties, such as length (Mant et al. 1988), secondary structures (Blondelle et al. 1995), sequence-specific correction factors and the ion-pairing separation mechanism (Krokhin et al. 2004; Krokhin 2006; Spicer et al. 2007). One such modern prediction algorithm considering a number of these parameters is the sequence-specific retention calculator (SSRCalc, publicly available at http://hs2.proteome.ca/SSRCalc/SSRCalc33B.html) (Krokhin 2006). Krokhin tested a set with ~2,000 tryptic peptides, which showed a linear correlation between peptide retention and hydrophobicity with an R^2 value of ~0.96. In the study presented here, we have used the SSRCalc for prediction of the retention times.

Often, one-dimensional peptide separation is not sufficient to resolve complex biological samples for MS analyses. Hence, an additional peptide separation step prior to RPLC is often introduced. Recently, IEF for peptide separation has become increasingly popular as a first dimension in a two-dimensional separation workflow (Cargile et al. 2004a, b; Eriksson et al. 2008; Fraterman et al. 2007; Heller et al. 2005a; Horth et al. 2006; Lengqvist et al. 2007). As the retention time, p*I* value prediction provides an independent means to support or disqualify peptide sequence assignments (Cargile et al. 2005; Krijgsveld et al. 2006; Heller et al. 2005b). Improved peptide p*I*-prediction algorithms for a narrow acidic p*I* range (~3.5–5.0) have recently been presented, e.g., by the groups of Bjellqvist and Stephenson (Cargile et al. 2008; Stephenson et al. 2006; Uhlén et al. 2006). Calculated p*I*s can be plotted and appropriate cutoff values applied to highlight less likely sequences (i.e., sequences showing outlier p*I* values) to curate protein identification result lists. However, an outlier does not necessarily have to be due to an erroneous match, but may indicate the presence of a PTM.

Biological and artifactual peptide modifications can shift both peptide p*I* and retention time compared to the values predicted for the native, non-modified sequence by introducing or neutralizing charged groups. Recently, a novel database search engine, the Paragon algorithm, was presented, which allows open-modification searches (in the default setting 154 Unimod modifications (Shilov et al. 2007). When identified peptide sequences from these open-modification searches are p*I* predicted and plotted, trends for the observed modifications on peptide p*I* can be derived. In the present study, general effects of a number of common as well as less common modifications and amino acid substitutions are described in terms of both observed p*I* and retention time shifts. Further, the degree of variation in IPG-IEF/RPLC separations is described using several complex biological samples.

Materials and methods

For the study of peptide p*I* shifts, LC–MS/MS data from IPG-IEF separated complex protein digests originating from two different human cancer cell lines were used: microsomal protein fractions (2 + 2) from the doxorubicin-sensitive/resistant lung cancer cell line pair H69 and H69AR. Biological aspects of the data sets used is presented elsewhere (Eriksson et al. 2008) and will not be discussed further in this study, which concerns only the behavior of peptides (detected peptide sequences) in IPG-IEF separations. For the study of retention time shifts, LC–MS/MS data from protein digests originating from dog plasma samples and human liver microsome samples were used.

The data from the cell lines H69 and H69AR have been submitted to the PRIDE database, accession number 13080 (http://www.ebi.ac.uk/pride/).

Workflow for peptide p*I* data set (H69/H69AR)

Sample preparation

Microsomal preparations of H69 and H69AR were prepared as described in (Eriksson et al. 2008). Following delipidation (Wessel and Flugge 1984), reduction and alkylation (TCEP and 4-vinylpyridine, 4-VP) and digestion (sequencing grade trypsin, Promega Corp, Madison, WI, USA), samples were iTRAQ labeled and pooled (2 H69 + 2 H69AR). Before IEF separation, excess reagent and detergent were removed (SCX-cartridge, Strata-XC, Phenomenex, Torrance, CA, USA) and the sample was dried in a Speedvac and finally solubilized in 8 M urea.

Isoelectric focusing of peptides

Tryptic peptide samples were dissolved in 225 μL 8 M urea. Narrow-range IPG strips for peptide focusing (pH 3.4–4.8 or 3.7–4.9, 24 cm long) together with dry sample application gels (42 × 5×1 mm) were kindly supplied by GE Healthcare Bio-Sciences AB, Uppsala, Sweden. The application gels were rehydrated in sample overnight, while the strips were rehydrated overnight in 8 M urea and 1% Pharmalyte^{TM} 2.5–5 (GE Healthcare Bio-Sciences AB, Uppsala, Sweden). The IPG strips were put in the focusing tray and the application gels containing the samples were placed on the anodic end of the IPG strips with filter paper between the application gels and the electrodes. The strips were covered with mineral oil and the focusing was performed on an Ettan^{TM} IPGphor^{TM} (GE Healthcare Bio-Sciences AB, Uppsala, Sweden) until 100 kVh had been reached. The strips were removed and hastily cut in 0.5 cm pieces (manually) and stored in −20°C until further use. Subsequently, extraction was performed by incubating the pieces in 50–100 μL milliQ-grade water for 1 h and repeated twice. In the pilot experiment, the peptides were eluted using a robotic device, without cutting the strips (performed at GE Healthcare Bio-Sciences AB, Uppsala, Sweden on prototype instrumentation). The extracts were dried using a Speedvac. Before LC separation, the fractions were dissolved in 15–20 μL milliQ-grade water with 0.05% TFA.

NanoLC analyses: peptide p*I* data set

NanoLC separation was performed on an Ultimate 3000 LC system (Dionex/LC Packings, Sunnyvale, CA, USA). Aliquots from chosen IPG fractions were applied using μL pick-up. As loading solvent as well as transport liquid, 0.05% heptafluorobutyric acid (HFBA) was used. Samples were first loaded on a 200 μm × 5 mm PS-DVB monolithic trap cartridge (Dionex/LC Packings, Sunnyvale, CA, USA) for desalting and concentration and then back-eluted onto the analytical monolithic column, PS-DVB 200 μm (Dionex/LC Packings, Sunnyvale, CA, USA). The flow rate was set to 1.5 μL/min. Solvent A was H_20/ACN/TFA (97:3:0.05 v/v/v) and solvent B was H_20/ACN/TFA (50:50:0.04 v/v/v). Peptides were separated using the following gradient: 0–8 min 0% B, 8–9 min 0–15% B, 9–39 min 15–95% B, 39–45 min 95% B, 45–55 min 0% B. The Probot fraction collector (Dionex/LC Packings, Sunnyvale, CA, USA) was set to collect fractions every 6 s between 10 and 25 min onto a blank MALDI target plate (Applied Biosystems). The eluate was mixed 1:1 (v/v) post column with 7 mg/mL CHCA (Bio-Rad Laboratories, Hercules, CA, USA) in 70% acetonitrile before being spotted onto the MALDI target.

MALDI MS analyses

MALDI analyses were performed on a 4800 MALDI TOF/TOF instrument (Applied Biosystems, Framingham, MA, USA). The instrument was operated in positive ion mode and externally calibrated in the peptide range using a mass calibration standard kit (Applied Biosystems). The mass spectrometer was set to perform data acquisition in the mass range of 700–4,000 *m*/*z*. In each MS spectrum, a maximum of 10–15 peptides were chosen (only peptides above a set *S*/*N* threshold, usually 80–100) for fragmentation, starting with the strongest peptide.

Data analysis

Peptide identification was carried out using the Paragon algorithm (Shilov et al. 2007) in the ProteinPilot 2.0 software package (Applied Biosystems, Foster City, CA, USA). The searches were performed against the IPI database (build 3.36, update 13 November 2007) or NCBI non-redundant (update 23 June 2007), limited to human sequences. Default settings for a 4800 instrument were used (i.e., no manual settings for mass tolerance was given). False discovery rates (FDR) were estimated by searching the data against a database consisting of both forward and reversed sequences. FDR was then calculated by the following formula: (2 × reverse hits/total number of hits) × 100. For the microsomal preparation data sets (H69 and H69AR), the FDR was 1%. The peptide p*I* values were predicted using an algorithm kindly provided by Stephenson et al. (2006). Only peptides assigned a high confidence score (≥99% confidence, ProteinPilot unused score; 2.0, 5,125 and 5,148 peptides for run 1 and 2, respectively) were used in the presented study. Average p*I* values and standard deviations for each IEF fraction were calculated after removing 10% of the most extreme p*I* values.

Workflow for retention time data sets (dog plasma and human liver microsomes)

Sample preparation

Two different samples were used for the investigations of retention time effects. One was a dog plasma sample consisting of the supernatant from acetonitrile precipitated plasma (according to Kay et al. 2008). The supernatant protein digest was separated by IPG-IEF and individual fractions were analyzed by nanoLC–MS.

The second sample was a commercial human liver microsome (Supersome) preparation. This sample was processed using the filter-aided sample preparation (FASP) methodology (Wisniewski et al. 2009). Sample processing is described for both samples below.

NanoLC separations: dog plasma (60 and 150 min)

Dog plasma was precipitated essentially as described (Kay et al. 2008). Briefly, 200 μL of dog plasma was diluted with 400 μL of water. Tubes were vortexed before adding 900 μL acetonitrile and placed in a sonicator bath for 10 min. Then tubes were vortexed once and placed in the sonicator bath for an additional 10 min. Samples were centrifuged at 12,000×*g* for 10 min and the supernatant was collected. This yields approximately 300 μg of protein in the supernatant fraction. The supernatant was evaporated to dryness and the proteins digested using trypsin according to the urea–ProteasMax combination protocol recommended by the manufacturer (Promega Corp.). Essentially, proteins were dissolved in 15μL 8 M urea and 20 μL 0.1% ProteasMax stock solution was added. After vortexing to dissolve the protein pellet, 58.5 μL of 50 mM ammonium bicarbonate was added. After reduction (5 mM DTT) and alkylation (15 mM iodoacetamide), trypsin digestion was carried out overnight. After digestion, the enzymatic reaction was stopped by acidification to pH 3 followed by SPE cleanup of the digests (Strata-X SPE cartridges, 30 mg bed, final elution 600 μL 70% acetonitrile, 0.1% formic acid). Eluates were evaporated to dryness before starting the peptide IPG-IEF separation protocol. Peptide IEF was carried out as described above but using wider pH range IPG strips (3–10 linear strips, GE Healthcare) than for the peptide IEF (H69/H69AR) data set. Peptide extraction was described using an extraction robot. IEF fractions were evaporated to dryness and reconstituted in 20 μL loading solvent (2% acetonitrile, 0.1% formic acid and 0.01% trifluoroacetic acid). Fractions were stored at −20°C until nanoLC analysis.

NanoLC separations: human liver microsomes (150 and 350 min)

Human UGT2B7 Supersomes^{TM} were obtained from Gentest, Woburn, MA, USA. The solution has 5.0 mg/mL of protein in 0.1 M Tris, pH 7.5. Tryptic digests were prepared using the FASP protocol published by the Matthias Mann group (Wisniewski et al. 2009) with minor modifications. Briefly, 24 μL of sample was mixed with 16 μL of 10% SDS giving 4% final SDS concentration. This mixture was boiled for 3 min at 95°C before centrifuging at 16,000×*g* for 5 min at 20°C. From the supernatant, 30 μL (corresponding to ∼90 μg of protein) was taken through the remaining steps of the FASP protocol [FASP I, see Supplementary Protocols online (Wisniewski et al. 2009)]. The 10 kDa molecular weight cutoff spin filters were used. A modification to the protocol was that the first digestion step was carried out using trypsin (in 40 µL 1 M urea) overnight instead of Lys-C. Further, the second digestion step was modified to mimic the urea–ProteasMax digestion described above. This was essentially done by adding 60 µL 0.1 M Tris HCl, pH 8.0 with no urea and 0.0833% ProteasMax to achieve a final detergent concentration of 0.05% in the resulting 100 µL reaction volume. This second digestion step was carried out for 4 h before stopping the reaction and SPE clean-up as above (30 mg Strata-X SPE cartridges). SPE eluates were evaporated (to dryness, reconstituted in 90 µL of loading solvent of 2% acetonitrile, 0.1% formic acid and 0.01% trifluoroacetic acid) and transferred to sample vials. The samples were stored at −20°C until nanoLC analysis. Sample loading to the nanoLC column was about 100–350 ng of peptide digest per injection as indicated in the figure legends.

NanoLC MS analyses (retention time data sets)

For the retention time analyses, the two types of samples described above (dog plasma and human liver microsomes) were analyzed using a NanoAcquity-Q-TOF Ultima Global instrument combination (both from Waters Corp., Milford, MA, USA) equipped with a nanoLockmass sprayer. The auxiliary pump of the NanoAcquity was used to (1) deliver a steady flow of calibrant (lockmass) solution (Glu-fibrinogen peptide solution, adjusted to give ~100 counts per second in continuum data acquisition) and (2) deliver a 50 nL/min flow of isopropanol to the analyte nanospray steel needle (30 µm i.d., 105 mm stainless steel emitters with 360 µm o.d. sleeves from Proxeon A/S, Odense, Denmark) through a stainless steel nanoTee (Waters Corp.). The isopropanol addition was done to assist and stabilize the nanoelectrospray as suggested (Anderson et al. 2009). A and B solvents were water and acetonitrile each with 0.1% of formic acid. The bridged ethyl hybrid C18 columns used (1.7 µm particles, 130 Å pore size) were either of 100 µm i.d. × 10 mm (for the 60 min analyses) or 75 µm i.d. × 250 mm (150 and 350 min analyses). The column was maintained at 35°C throughout. For efficient sample loading, a trap column configuration was used. Trapping was done at a 99.9% A-solvent concentration at a 5 µL/min flow rate for 3 min. The initial separation conditions were 99% A-solvent at a flow rate of 350 nL/min. A gradient was then run between 1 and 40% B followed by a ramp to 95% B, a hold step at 95% B and a re-equilibration step. These steps were adjusted to a 60, 150 or 350 min cycle time. Detailed conditions are available upon request. Data were acquired in centroid mode at a working resolution of the instrument of ~8,300 (FWHM as determined for the 785.8426 *m/z* glu-fib ion).

Retention time determination

The peak retention times were obtained either manually (using the MassLynx software to display extracted ion chromatograms) or by using the Positive software package to build QuanLynx methods for each peptide ion to be determined (the QuanLynx method was set to use the ApexTrack algorithm). Additional data processing was done in Excel.

Retention time prediction

In order to use the SSRCalc online tool to predict retention times (available at http://hs2.proteome.ca/SSRCalc/SSRCalc33B.html), one has to trim the equation according to the LC system and gradient used. This can be done in different ways, where one is to use a digest of a known protein (e.g., BSA), extract the experimental retention times of identified peptides and plot these against the hydrophobicities of the peptides (calculated in the SSRCalc). From this dependence, $\mathrm{RT} = A + B \times (\mathrm{HP})$, the A and B parameters are determined where A is the gradient delay time and B is a value related to the slope of the acetonitrile gradient. The A and B parameters are then used in the SSRCalc program, together with information about column properties, to predict retention times of other identified peptides in any runs under the same conditions. Using 31 identified BSA peptides (Mascot scores above 30), we determined the A and B parameters for our system. Plotting observed retention times against predicted, we could see a linear correlation with an R^2 value of 0.93 (data not shown), indicating adequate predicting of the SSRCalc algorithm.

Peptide mass queries of the human SwissProt protein database

From a FASTA-file holding the protein sequences in the SwissProt database (release 57.14 of 09 February 2010), human protein sequences were extracted to generate another FASTA-file. This file was queried using the msInspect software tool (Bellew et al. 2006) from the Computational Proteomics Laboratory (CPL) at Fred Hutchinson Cancer Research Center (http://proteomics.fhcrc.org/CPL/msinspect/index.html) installed on a local computer. For the data shown in Fig. 4, theoretical exact mass values for nine selected peptide sequences (CDVDIR, DLTDYLMK, DSYVGDEAQSK, DSYVGDEAQSKR, QEYDESGPSIVHR, SYELPDGQVITIGNER, DLYANT VLSGGTTMYPGIADR and LCYVALDFEQEMATAASS

SSLEK) were used. These peptides were all assigned to the UGT2B7 protein [EC = 2.4.1.17, identified from the human liver microsome sample using Mascot searching (one missed cleavage, database: SwissProt release 57.14, constant modification: carbamidomethyl Cys, variable modification: oxidation of Met, precursor tolerance: 0.1 Da, fragment ion tolerance: 0.2 Da)]. The nine peptides ranged from 4.00 to 4.65 in p*I* and from 776.3487 to 2549.1665 in mass. For each peptide sequence, theoretical mass values were input to the FASTA-query tool of msInspect and peptide sequences were extracted with a mass tolerance of ±50 parts per million (ppm) from the Swissprot human database. Using the Stephenson p*I*-prediction algorithm and the SSRCalc tool, peptide p*I* and retention time values were calculated. Retention time values were calculated for the 350 min separation (Fig. 2) using 42 peptides from a BSA digest for calibration of the retention time scale ($R^2 = 0.929$). The model used in the SSRCalc tool was for TFA as an ion-pairing agent and a 100 Å C18 stationary phase material.

Results

Experimental accuracy of peptide IPG-IEF and nanoLC–MS separations

Proteomic analyses often include multidimensional separations. In the study presented here, the experimental precision has been investigated for narrow-range IEF (pH range 3.4–4.8) using 24 cm IPG strips and for nanoLC–MS analysis using a Waters nanoAcquity system.

To investigate the correlation between predicted and observed peptide p*I* value, peptide identification data from two separate analyses of a human microsome sample were used as described below. In Fig. 1 is plotted the average p*I* values (determined for the bulk of the peptides identified in each fraction) together with the standard deviation for ten IEF fractions. These fractions were obtained from the basic end of the narrow-range IEF-IPG strips, as can be seen from the p*I* scale (*y*-axis). The values obtained from two replicates of the entire workflow are shown, i.e., the processing of two aliquots of microsomal preparations of the small cell lung cancer cell lines H69/H69AR (denoted as run 1 and 2). The peptide p*I* standard deviations are all below 0.055 pH units (Fig. 1). In the calculation of average fraction p*I* values, the 10% most extreme p*I* values were excluded to remove the influence of PTM-induced p*I* shifts and erroneous peptide assignments. These average values were then used in the calculation of PTM-induced shifts for individual peptides (see below). Further, the high reproducibility of the IEF separation method is indicated by the low deviation in average fraction p*I* values between the two runs (<0.1 pH-units, see Fig. 1).

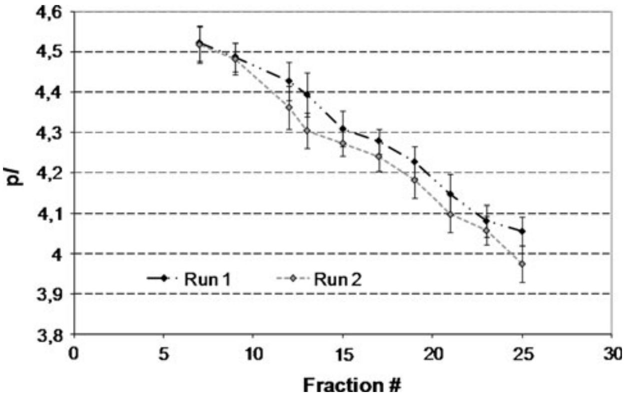


Fig. 1 Calculated average p*I* values (extreme values excluded) in ten analyzed IEF fractions from two replicate runs. The standard deviations in each fraction are also shown

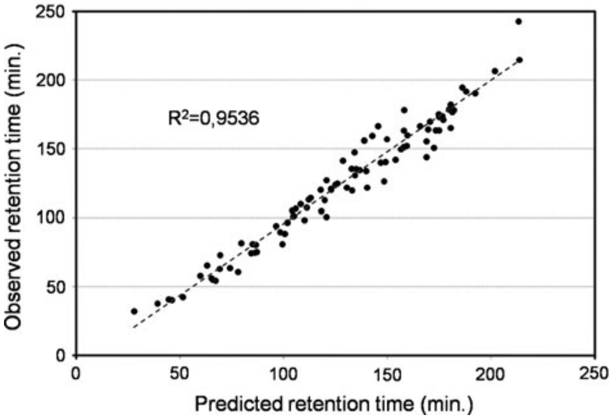


Fig. 2 Predicted retention times for the 350 min gradient were calculated for 100 peptides identified (Mascot score >40) from the liver microsome digest. Observed (*y*-axis) versus predicted retention times (*x*-axis) are plotted

To investigate the correlation between theoretically predicted and experimentally observed peptide retention time, a non-fractionated highly complex proteomic sample, namely a liver microsome protein digest, was used. Predicted retention times were calculated for 100 peptides identified with high confidence from the liver microsome digest (Mascot score >40, 350 min nanoLC gradient). Figure 2 shows the plotted differences between observed (*y*-axis) and predicted retention times (*x*-axis). A linear trend line could be fitted with an R^2 value of 0.9536 indicating a good correlation. In absolute terms, the differences between observed and predicted retention times ranged from −25.3 to +29.2 min (standard deviation: 10.15 min).

To evaluate the experimental retention time reproducibility in our analysis, the observed retention times for a

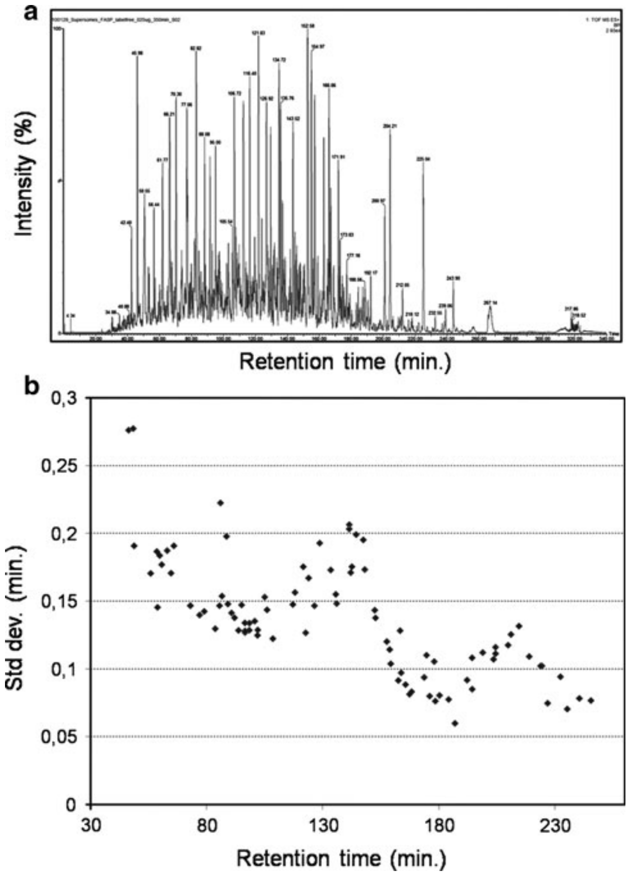


Fig. 3 **a** A representative BPI chromatogram from a 350 min nanoLC analysis of ~0.35 µg liver microsome protein digest. **b** Individual standard deviations for 90 randomly selected peptides covering the span of the gradient in eight consecutive 350 min nanoLC runs

number of peptides from bovine serum albumin (BSA) were inspected in six LC–MS replicate runs for a 150 min nanoLC gradient (back-to-back injections, see Table 1 in supplementary data). Peptides identified with a Mascot score above 30 were chosen and the shifts were small; calculated on 24 peptides, the mean standard deviation was 0.105 min.

The experimental precision in retention time was also investigated for the more complex microsome digest sample. To assess reproducibility, eight back-to-back injections of ~0.35 µg of protein digest were analyzed using a 350 min nanoLC gradient on a 75 µm × 250 mm C18 BEH-column. A representative BPI chromatogram is shown in Fig. 3a. Though not baseline separated, the sharpness of the peaks and their even distribution indicate a satisfactory separation of this highly complex non-fractionated sample. As a measure of retention time reproducibility, 90 of the eluting peptides were randomly selected to cover the 46–245 min elution time interval. The individual standard deviations for this peptide set are plotted in Fig. 3b as a function of retention time. Standard deviations ranged from 0.059 to 0.270 min. We have observed similar low retention time deviations over extended analysis of real samples. Shown in Supplementary Fig. 1 is plotted the retention time for the LVNELTEFAK peptide of a BSA digest used as a quality control sample and injected repeatedly over a 3-day sample analysis period. The peptide elution window is from ~30 to ~250 min, i.e., covering about 220 min elution time in total. Peak widths are 1.0–1.3 min at baseline, as determined by manual inspection.

Information content of observed peptide mass values in combination with p*I* and Rt prediction

Based on the experimentally observed accuracy from our experiments (Figs. 1, 2, 3), the added value of using predicted p*I* and retention time constraints was investigated for human proteomic analyses. The information content was estimated by plotting nine theoretical peptide masses from the randomly selected UGT2B7 protein against the entire human tryptic peptidome (one missed cleavage site, SwissProt database containing 21,273 protein sequences) (Fig. 4). For a ±50 ppm mass constraint, the number of possible peptide matches range from ~175 to 450 (Fig. 4a, each constraint is indicated in the figure legend). The number of possible matches is higher for lower peptide masses on the left-hand side of the plot than for higher mass peptides. In Fig. 4b shows the plot of the number of possible matches after applying each of three further constraints as indicated (±10 ppm mass accuracy, ±0.2 p*I* unit accuracy or ±20 min difference in predicted retention time). As can be seen, all of these constraints significantly reduce the number of possible peptide matches by about 60–80% or more (Fig. 4). There is a clear trend that the retention time constraint is the least discriminatory constraint (Fig. 4a). For the retention time filter, the ±20 min constraint was chosen as it is ±2*, the observed standard deviation for retention time prediction (10.15 min, see above).

Further, the cumulative effect of applying these constraints in p*I* and retention time was tested. The number of peptide matches for each mass value after consecutively applying the 10 ppm, 0.2 p*I* and 20 min retention time filters are plotted in Fig. 4b. This resulted in two to five possible peptides sharing the characteristics of the true identification for all peptides tested except for the lowest mass peptide (*m/z* 776, Fig. 4). For 1/9 peptides, the true positive hit was the only hit. For modern high mass accuracy machines such as OrbiTrap, FTICR or QTOF instruments, mass accuracy < ±5 ppm is readily achievable. Plotted in Fig. 4b is the number of peptide matches after applying a final (i.e., after p*I* and retention time filtering) ±5 ppm mass constraint (white bars).

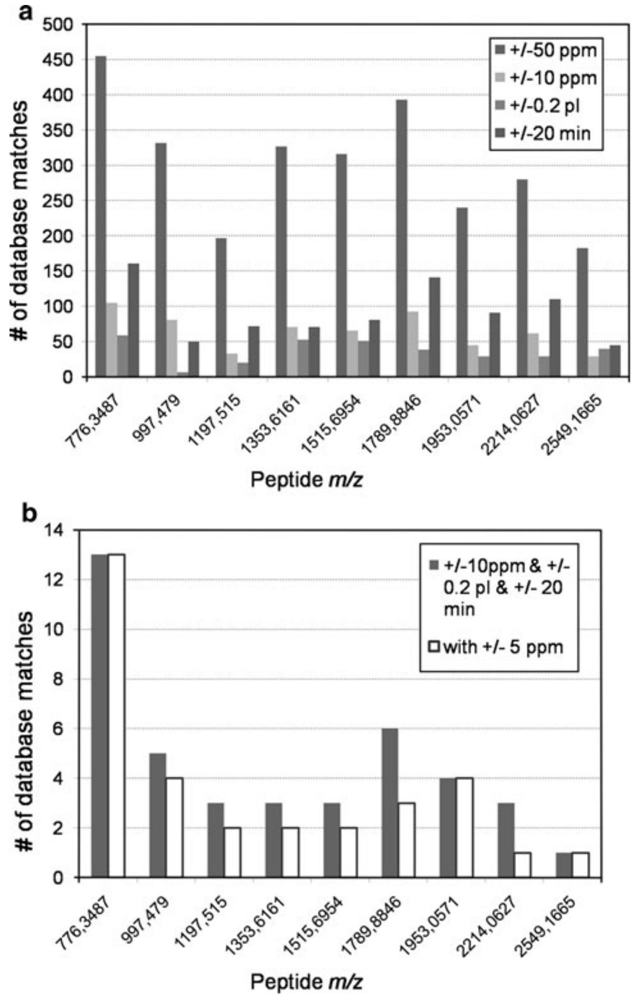


Fig. 4 **a** The number of shared peptide sequences from a 50 ppm mass window around each of the nine selected peptides from the UGT2B7 protein identified from the liver microsome sample. Further, the added effect of applying either of three constraints (±10 ppm, ±0.2 p*I* or ±10 min) is plotted. **b** The cumulative effect of consecutively applying the mass, p*I* and retention time filters is plotted

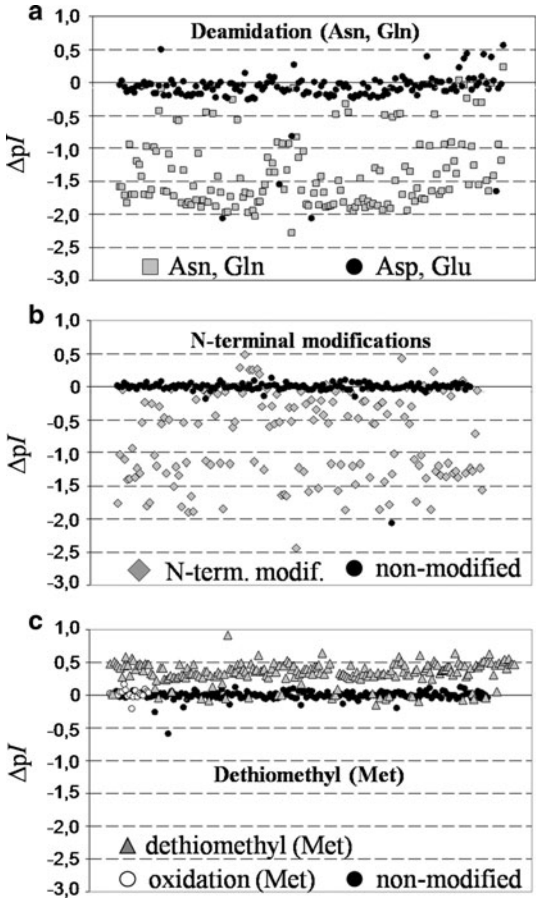


Fig. 5 **a** Peptides denoted as having deamidated Asn or Gln were selected and an in-house perl-script was used to convert Asn and Gln to Asp and Glu, respectively. The calculated p*I* values before (*gray squares*) and after the conversion (*black dots*). **b** Calculated p*I* values of peptides with an N-terminal modification (*gray diamonds*), compared to non-modified peptides (*black dots*). **c** Calculated p*I* values of peptides with oxidized Met (*white spots*), dethiomethyl Met (*gray triangles*) and control (*black dots*)

p*I* shifts induced by common amino acid modifications

Drawing on the high experimental accuracy of the narrow-range peptide IPG-IEF method, experimental p*I* values were determined for high confidence peptide identifications flagged as PTMs in an open-modification search (peptide confidence >99% in a Paragon search for the cancer cell microsome digest, see Fig. 1). To investigate the effect of amino acid modifications on the experimentally observed peptide p*I*, average fraction p*I*s (determined and plotted in Fig. 1) were subtracted from predicted p*I* values for each detected modified peptide. As the peptide p*I* calculation is done for non-modified sequences, peptides which have a shifted p*I* due to a post-translational modification will show up as outliers. Further, the magnitude of the shift can be assigned for each type of modification. Shifts were plotted for the most frequently observed modifications (deamidations of Asn and Gln, N-terminal modifications and dethiomethyl on Met) (Fig. 5). It was previously observed that deamidation of Asn and Gln (i.e., conversion to Asp and Glu, respectively) renders peptides significantly more acidic and they thus show up as outliers in p*I* plots (Lengqvist et al. 2007). In this study, an in-house script was used to compensate for this by making the Asn → Asp and Gln → Glu exchanges before plotting p*I* values. In Fig. 5a, high confidence peptides assigned as deamidated are plotted before and after this correction (denoted as gray squares and black dots, respectively). A clear tendency showing a shift of ~1.5 p*I* units can be observed. This shift is large compared to the experimental standard deviation (<0.05 p*I* units, see Fig. 1) and corresponds to the introduction of an additional carboxylic acid group in the peptide. Modifications affecting the N-terminal amine (neutralization of the basic group) also drastically affect peptide p*I*. Figure 5b shows a collective plot for all N-terminal acetylations and Gln → pyro-Glu conversions detected. As can be seen, the shifts observed range from ~0.5 to 2.0 p*I* units. This observation is consistent with a reduction in the number of dissociating amino acid residues.

An unexpected shift toward more basic p*I* values (+~0.4 p*I* units) was observed for peptides denoted as having dethiomethyl Met modifications, shown by gray triangles in Fig. 5c. This modification is present at relatively high levels: 1.8 and 2.3% of all peptides in run 1 and 2, respectively. It is clear that dethiomethyl peptides group separately from either control (non-modified) peptides or peptides with oxidized Met residues (indicated by black dots and white dots, respectively in Fig. 5c).

To assess the reproducibility in the detection of peptide p*I* shifts, a smaller high confidence peptide set was investigated. This consisted of N-terminally modified peptides detected in both replicate runs. Figure 6 shows the plot of the determined p*I* shift for carbamidomethylated N-termini (a), acetylated N-termini (b) and Gln to pyro-Glu conversions (c).The differences between the IPG-IEF run 1 and run 2 measurements are consistently ~0.02 p*I* units or below (Fig. 6a–c), further substantiating the good reproducibility between experiments of the IPG-IEF method.

RPLC retention time shifts induced by common amino acid modifications

The modifications from previous work that were known to alter the retention time (Zybailov et al. 2009) and modifications shown to alter the p*I* value were analyzed for retention time effects. Peptides for which both the unmodified and modified forms were present were extracted from the data set. Figure 7 shows the observed retention time shifts both as a population (a) and for each individual peptide (b–e). Oxidation of methionine (Fig. 7b) shortened the retention time of the modified peptides with 0.7–3.0 min (15 peptides; mean value 1.3 min). Formylation at the N-terminus increased the retention time with 0.6–5.9 min (ten peptide pairs; mean value 2.8 min) (Fig. 7c). Conversion of glutamate to pyro-glutamate prolonged the retention time with 2.5–7.1 min (4 peptide pairs; mean value 3.9 min) (Fig. 7d). Conversion of glutamine to pyro-glutamate prolonged the retention time with 0–5 min (4 peptide pairs; mean value 2.5 min) (Fig. 7e). It should be noted that these assignments were made from 60 min nanoLC runs of individual fractions of IEF fractionated dog plasma (see "Materials and methods" section).

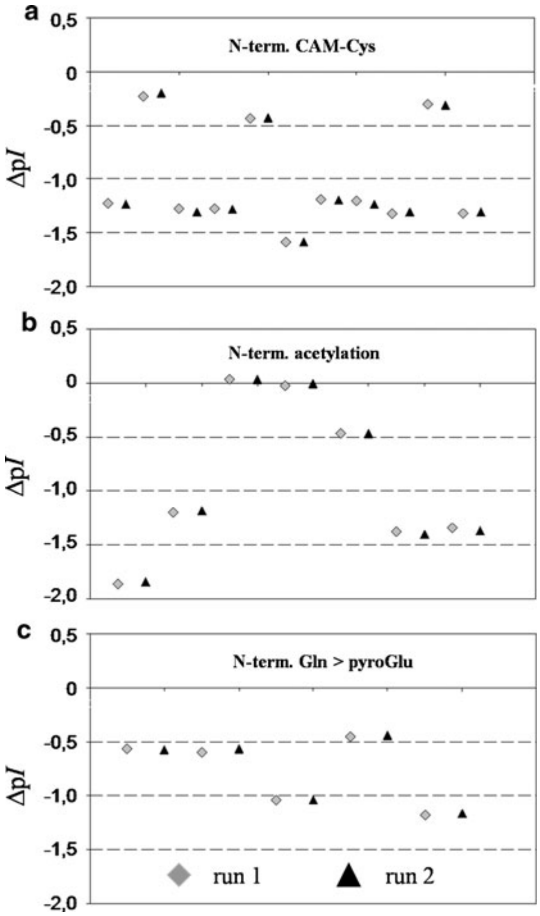


Fig. 6 Peptides with a modification at the N-terminal showed a shift in the p*I* compared to the fraction average. **a** CAM-cysteine (N-terminal), **b** acetylation (N-terminal) and **c** Gln > pyro-Glu (N-terminal). *Gray diamonds* indicate replicate run 1 and *black triangles* replicate run 2

Detection of amino acid substitutions

One application that benefits from high confidence in peptide sequence assignment is the identification of amino acid substitutions. One such assignment based on corroborative experimental evidence is outlined below. The Paragon search algorithm can assign not only modification, but can also suggest amino acid substitutions (Shilov et al. 2007). Figure 8 shows the MS/MS spectrum for a peptide denoted as a Glu → Gln conversion. The peptide sequence assigned (99% confidence) was FESPEVAER (theoretical *m/z* 1207.61). However, as the search algorithm considers the amino acid substitution, the actual peptide detected has the sequence FQSPEVAER (*m/z* 1206.63, a Glu → Gln conversion is associated with a −0.98 Da mass shift). This sequence has a calculated p*I* of 4.50, which compares favorably with the fraction p*I* (4.46, SD 0.046). In contrast, the FESPEVAER sequence has a p*I* of 4.18. On close inspection, the fragment ion series supports the detection of

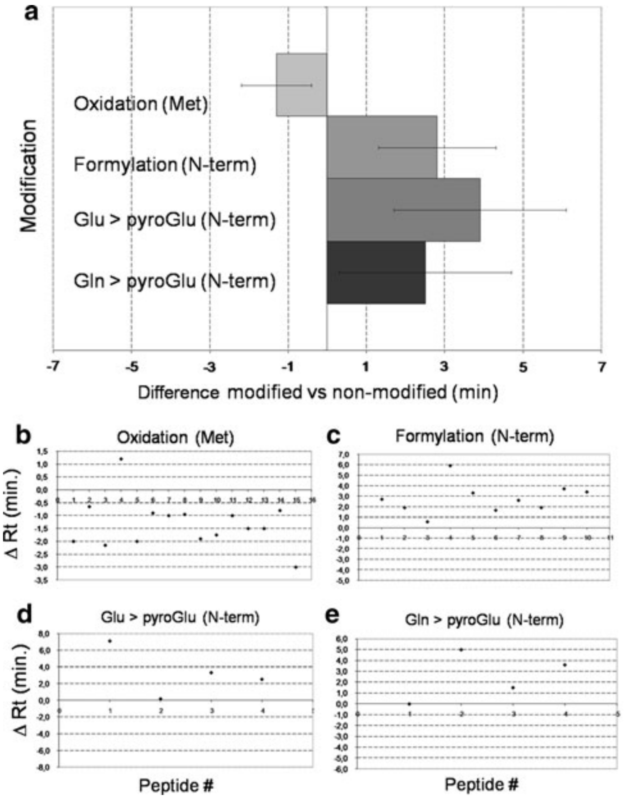


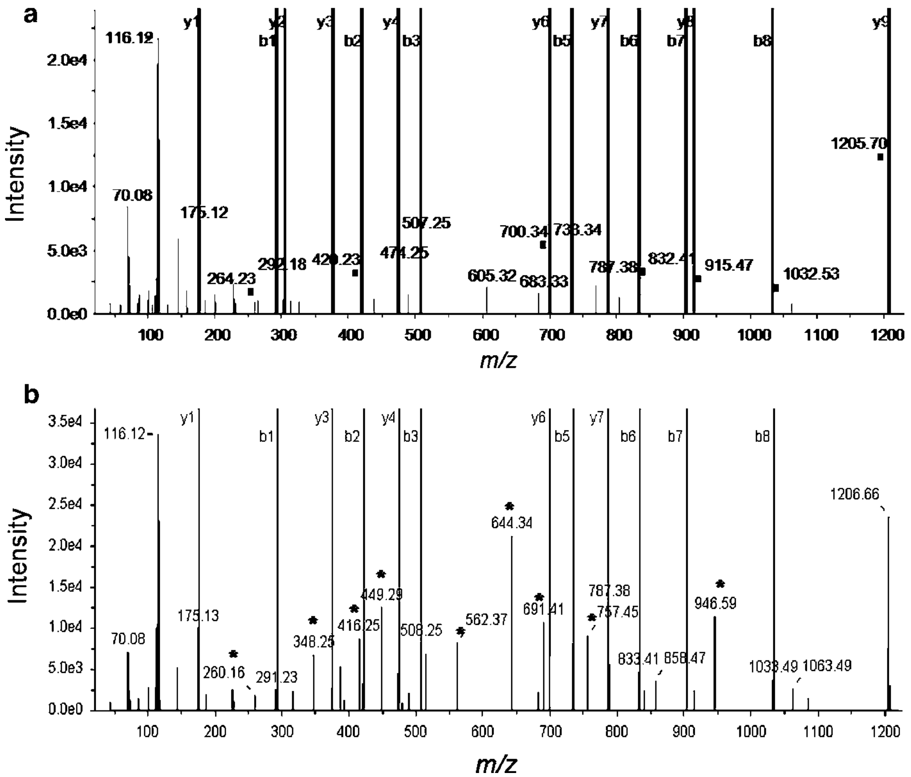
Fig. 7 **a** Average differences in elution times in the nanoLC (60 min run, dog plasma) analysis for four chosen modifications with the standard deviations. **b**–**e** Each modification and the shift in retention time for peptides with both modified and non-modified form present in the analysis

the Gln containing sequence (see Fig. 8a). The MS/MS spectrum is shown on the left and on the right are theoretical b- and y-ions with ions not detected in parenthesis. If the mass shift is localized to the second position E-residue, normal b^1 and y^1-y^7 ions are expected. This is indeed observed albeit with the absence of the b^4, b^9 and y^5 ions (Fig. 8a). Interestingly, the native sequence (FESPEVAER) was detected, but in another IEF fraction (99% confidence). This fraction had an average p*I* of 4.092 (SD 0.055), which is in good agreement with the sequence p*I* of 4.18. The MS/MS spectrum of the *m/z* 1207.61 precursor is given in Fig. 8b. Comparing the two right-hand tables in Fig. 8, the fragment ions observed supports the E containing sequence. Fragment ions common to both sequences are shown in bold and specific fragment masses are underlined, while absent ions are in parenthesis. Notably, the defining y^8-ion (*m/z* 915 and 916 for Q and E, respectively) was present in both cases. The MS/MS spectrum of FESPEVAER also contains another fragment ion series. This series originates from the co-incidental fragmentation of another precursor ion at an *m/z* of 1,205 and is indicated by asterisks in Fig. 8b. The FESPEVAER sequence was also detected a third time with high (99%) confidence. This was as the methylesterified peptide (on the first E-residue carboxyl) having a precursor *m/z* value of 1221.63. This identification was from the same fraction as the FQSPEVAER peptide (average p*I* 4.46). Indeed, substituting the Glu for a non-charged amino acid (Gly or Val, F[G/V]SPEVAER) gives a predicted p*I* of 4.50 in close agreement with fraction p*I* (4.46).

Discussion

It is clear that many of the peptide species observed in LC–MS shotgun-based proteomic analyses are due to sample processing artifacts (Ahrne et al. 2009; Nielsen et al. 2006; Xie et al. 2009). The most abundant of these are listed in Table 1. From an analytical chemistry point of view, measuring all existing forms of a peptide is necessary for accurate quantification. Furthermore, in a complex proteome experiment, the theoretical search space of both modified and non-modified peptides is immense. Here we show that combined narrow-range IPG-IEF separations with nanoLC–MS detection are highly informative in limiting the number of potential peptide sequences associated with an observed mass value. In the analysis of selected UGT2B7 peptides, two to five candidate sequences are isobaric to the true peptide in terms of mass, p*I* and retention time (Fig. 4). Considering only a limited number for potential modifications (5), the number of tryptic peptide candidates associated with a single mass value will increase tenfold even at a ±5 ppm mass accuracy (Ahrne et al. 2009).We have shown that all three constraints (i.e., mass, p*I* or retention time) can independently reduce the number of candidate sequences for a given mass value. For an open-modification search strategy, these constraints can be used to filter the results list for increased confidence. As shown in Table 1, modifications exists that are silent (i.e., do not induce a shift) in terms of p*I* and retention time. Other modifications strongly shift either one or both of these parameters. That these parameters can be used for informative queries of experimental data was shown for the amino acid substitution example herein, keeping in mind that the combination of accurate mass and p*I* information is not sufficient to identify any given peptide from a complex (mammalian) proteome (Cargile and Stephenson 2004).

Additional sample pre-fractionation information can also be utilized to identify artifacts caused by ionization in MS analyses. As an example, instrument conditions may induce, e.g., oxidation and fragmentation. Berg et al. 2006 could differentiate between two oxidized forms of the same peptide (i.e., two isobaric mass values) in LC–MS data as one occurred in the same scan as the non-oxidized peptide and the other eluted minutes earlier (see Table 1). Thus, one oxidation event took place in real time (most likely in



Residue	b-ions	y-ions
F	292,18	1206,62
Q	420,24	915,45
S	507,27	787,39
P	(604,32)	700,36
E	733,36	(603,31)
V	832,43	474,27
A	903,47	375,20
E	1032,51	304,16
R	(1188,61)	175,12

Residue	b-ions	y-ions
F	292,18	1207,61
E	421,22	916,44
S	508,25	787,40
P	605,31	700,36
E	734,35	(603,31)
V	833,42	474,27
A	904,45	375,20
E	1033,50	(304,16)
R	(1189,60)	175,12

Fig. 8 Identification of amino acid substitutions. Shown are MS/MS spectra of peptides identified as **a** FQSPEVAER and **b** FESPEVAER. Sequences along with *b*- and *y*-ions are given with non-observed ions in *parenthesis*. Ions that are specific for each sequence are *underlined*

the electrospray process). while the other occurred prior to LC analysis. Similarly, we have observed outliers in our p*I* plots that arise from the sequencing of shorter versions of larger peptides. As they are present at the same LC elution time (LC–MALDI spot or ESI scan), the shorter peptides are most likely formed by in-source fragmentation.

As shown in Table 1, other events besides modifications are present as peaks in the LC–MS data. These include amino acid substitutions as well as sodium and potassium adducts formation. It is important to consider that different modifications and substitutions may induce the same mass shift, e.g., for Ala → Ser conversion and oxidation of Met, His or Trp. These conversions are also non-distinguishable in terms of retention time and p*I*. Hence, high confidence identification of modified peptides from shotgun LC–MS/MS data should rely heavily on the MS and MS/MS data recorded. This includes verification of fragment ion coverage and inspection of the precursor ion mass spectrum to rule out co-fragmenting peptides as discussed (Fig. 8). It should be stressed that mass–p*I*-retention time filtering as discussed here should be used as filtering criteria and not confused with identification.

In this study, an amino acid substitution was detected (FESPEVAER to FQSPEVAER, i.e., Glu to Gln), which was further supported by the calculated p*I* and MS/MS spectrum (Fig. 8). This peptide identified the heterogeneous nuclear ribonucleoprotein M, hnRNP M (AC: P52272). This amino acid substitution has not been reported previously as a missense SNP (single nucleotide polymorphism) for the hnRNP M protein (Entrez SNP database at NCBI). However, on a genomic level the mutation appears plausible. The codons for Glu (GAA, GAG) can be converted to Gln codons (CAA, CAG) through a single G → C transversion. Such transversions occur readily for instance under conditions of oxidative stress [for mechanism see e.g., (Kino and Sugiyama 2001)]. The residue (position 700) is localized in an RNA recognition motif (exon 16 of the hnRNP M gene). The detection of both forms in the mass spectrometry data could indicate the presence of SNP heterozygosity. However, due to the possibility of artifactual deamidation of Gln, the presence of homozygous substitution cannot be ruled out, although it would have occurred prior to IEF separation. Hence, this example illustrates the potential use of p*I* data together with MS data to detect amino acid substitutions.

The narrow-range IPG-IEF format used also offers significant reduction of sample complexity. As shown by us and others, about 30% of all human peptides will cluster in the acidic p*I* range (3.5–5.0) (Eriksson et al. 2008; Cargile

Table 1 Common observations of modified, substituted and adducted amino acid residues in shotgun proteomics experiments and their prevalence as well as associated mass, p*I* and retention time shifts

Modification	Residue	Origin	Preval.	Observ.	Mass Δ	p*I* Δ	Rt Δ
Oxidation	M	Artifact	∼10%[a]	0.2%	+15.99	0	←
Deamidation	N, Q	Artifact	∼3%[a]	1.6%	+0.98	−∼1.5	±1[b]
Formylation	N-term	Artifact	∼3%[a]	0.03%	+27.99	–	→→
Formylation	S	Artifact	∼2%[a]	0.01%	+27.99	0	→
Dioxidation	M	Artifact	∼1%[a]	–	+31.99	0	←←
Gln → pyroGlu	N-term Q	Artifact	∼1%[a]	0.2%	−17.03	–	→→
Glu → pyroGlu	N-term E	Artifact	∼1%[a]	0.1%	−18.01	–	→→
Dethiomethyl	M	Artifact	Considerable	2.1%	−48.00	+∼0.4	?
Propionamide C	C	Artifact	∼1%[a]	–	+71.04	0	0
Acetylation	N-term	Biological	∼0.5%[a]	0.5%	+42.01	−	→→
Methylation	L	Biol./artifact	Considerable	–	+14.02	0	(→)
Methylation	D, E	Biol./artifact	Considerable	0.4%	+14.02	+	(→)
Glu → Gln	E	aa subst.	∼0.5%[a]	–	−0.98	+	0
Cationization (Na)	D, E	Artifact	∼0.5%[a]	–	+21.98	0	0
Formylation	T	Artifact	∼1%[a]	–	+27.99	0	→
Asp → Asn	D	aa subst.	∼0.5%[a]	–	−0.98	+	0
Carbamidomethyl	N-term	Artifact	Considerable	0.5%	+57.02	–	0
Ala → Ser	A	aa subst.	∼0.5%[a]	0.01%	+15.99	0	←
Oxidation	H, W	Artifact	∼0.5%[a]	0.4%	+15.99	0	←

Data compiled from this study and (Zybailov et al. 2009; Nielsen et al. 2006; Xie et al. 2009; Berg et al. 2006). Specific values are from our experiment ("observ." column)

[a] Zybailov et al. (2009)

[b] Xie et al. 2009; Berg et al. (2006)

The Rt shift to earlier or later elution is indicated by the arrows (← or → , respectively)

et al. 2004b). A theoretical tryptic digest of the 21,273 human proteins in the Swissprot database will, allowing for one missed cleavage site, result in a total of 2,045,133 peptides. Accordingly, about 600,000 of these will be observable using narrow-range IPG-IEF corresponding to close to 96% of all human proteins (Eriksson et al. 2008). The observed standard deviation in the present work was ≤0.055 p*I* units. This is very much in agreement with the accuracy reported for the peptide p*I*-prediction algorithm used here, 0.03 p*I* units (Cargile et al. 2008). Assuming an even distribution of peptides across the 3.5–5.0 range, ∼88,000 peptides will fall into each theoretical bin of ±2* the standard deviation (i.e., 0.22 p*I* units per bin, leading to 6.8 bins/narrow-range IEF experiment). However, the distribution is close to a normal distribution across this range, so over the 3.5–5.0 range some bins will be more heavily populated than others (Eriksson et al. 2008). This may warrant a guided pooling strategy for fractions along the length of the gradient as recognized by (Vaezzadeh et al. 2008). In terms of retention time precision, standard deviations of <0.3 min were observed for a highly complex peptide mixture (Fig. 3) and <0.1 min for single protein digests (Supplementary Table 1). Compared to the length of the peptide elution window, which is about 220 min for this 6-h LC run, the standard deviation is small (1.4 parts per thousand). The theoretical example outlined above can be expanded to include the retention time dimension. Dividing the 220-min peptide elution space into bins of ±2* standard deviation (1.2 min per bin) gives a total of ∼183 bins for the retention time space. Thus, in a combined IPG-IEF–nanoLC separation, an estimate of 480 peptides would populate each theoretical bin (88,000 peptides/183), limiting the search space considerably. This example highlights the added information content achievable from combining modern narrow-range IPG-IEF with high precision LC separations. However, here the experimentally observed standard deviation was used as the limit (i.e., ±2 × 0.3 min). The variation associated with predicting the retention time was much higher with a standard deviation of 10.15 min. This shows that retention time prediction lags behind experimental accuracy by a factor of about 10. Thus, though the predictive accuracy (i.e., *a* ± 20 min retention time filter) is informative today as shown here, refining retention time prediction algorithms is an important area of further research.

While the behavior of non-modified fully tryptic peptides can be predicted with high accuracy in terms of both p*I* (Cargile et al. 2008; Uhlén et al. 2006) and retention time (Krokhin 2006; Gilar et al. 2007), post-translationally modified peptides have been less explored. Here, several effects of individual amino acid modifications are described. Modifications that drastically alter the buffering power of a peptide by increasing or reducing the number of dissociating amino acids will show drastic effects on p*I*. In general, deamidations of Gln and Asn will cause a decrease in p*I* by about 1.5 p*I* units (Fig. 5). It will thus be possible to follow these relatively prevalent modifications to a considerable extent in narrow-range IPG-IEF data. We and others (Nielsen et al. 2006; Zybailov et al. 2009) have observed only minor changes in retention time for deamidations. However, Berg et al. 2006 observed a small shift of ∼1 min earlier elution. Similarly, Xie et al. (2009) observed two peaks, one eluting slightly before and one after the non-modified peak. These were identified as aspartic acid and isoaspartic acid forms for an Asn deamidation. Importantly, the different (aspartic/isoaspartic acid) forms are only distinguishable in the LC dimension, highlighting the need for combining information from both sources. Interestingly, a novel effect was shown for peptides that were detected as dethiomethyl Met-modified (−48.00 Da). Although the mechanism behind this is unclear, the general trend for this modification was to induce a shift of about +0.25 p*I* units (Fig. 5c). Though present in our data to a relatively large extent, the results of Nielsen et al. (2006) suggest that the detection of −48.00 mass shifted peptides may be sample dependent. Recent work on extending the functionality of p*I*-prediction algorithms to include PTMs (Gauci et al. 2008) has been presented. Future efforts to improve such algorithms and open-modification search engines can only benefit from the degree of accuracy possible using narrow-range IPG-IEF.

Future aspects

We postulate that mass–p*I*-retention time filtering can be used as a strategy for database reduction prior to an open-modification search that is independent of MS/MS data acquisition (Ahrne et al. 2009). In such a strategy, a candidate list of protein accessions would be obtained by calculating which candidate peptide sequences fit in the recorded and calculated mass, p*I* and retention time windows. Such an approach could also take into consideration the most common modifications and amino acid alterations (Table 1). As modifications do not occur at equal rates, such an implementation should employ some form of differential weighting system. One could envisage a scoring system based on modification prevalence. For example, an oxidation of Met is more likely for a +15.99 Da mass shift than an Ala → Ser substitution. Similarly, a scoring system could be introduced for matches to modified variants of already detected peptides. This is similar to the Modifi-Comb workflow where mass and retention time-shifted so-called “dependent” (variant) peptides are assigned for non-modified “base” peptides based on similarity of MS/MS spectra and accurate precursor mass measurements (Savitski et al. 2006). Although this limits the sensitivity of such an approach since only already identified peptides are considered, modified versions of high abundance peptides are a major part of the modified peptides observed (Xie et al. 2009; Nielsen et al. 2006; Ahrne et al. 2009). Using the outlier peptides based on pI and retention time shifts could mark a peptide cohort in a shotgun proteomic experiment for more rigorous open-modification search algorithms. Prior public domain knowledge on particular peptide modifications detected before in other experiments, i.e., present in public data repositories such as Peptide Atlas (Desiere et al. 2005) or the GPM (Craig et al. 2004) database, could also utilized in such algorithms.

The high degree of accuracy observed for both narrow-range IPG-IEF and nanoLC–MS methods clearly facilitates re-acquisition of previously observed peptides for data-dependent analysis. If these procedures are standardized, it could be beneficial to keep track of which mass–p*I*-retention time features are detected to form a database to facilitate the use of this information in proteome-wide experiments. This is similar to the accurate mass and time (AMT) tag concept described by the Richard Smith Laboratory (see e.g., Norbeck et al. 2005). In this approach, the sample is initially exhaustively characterized using LC–MS/MS and database searching. The recorded retention time of each identified peptide together with high mass accuracy measurements is then used to re-detect each peptide in subsequent experiments. It has been observed that the usefulness of the AMT approach increases with increased peak capacity of the RPLC separation and increased reproducibility of peptide elution times (Norbeck et al. 2005). Similar to the AMT method outlined above, a strategy incorporating peptide IEF information and accurate mass measurements has also been described (Cargile and Stephenson 2004). Recently, spectral library searches for peptide identifications have been lifted up as a promising identification method (Lam and Aebersold 2010). Similarly, pre-fractionation data could be used to limit the search space, leading to increased accuracy and computational speed of analysis. This work can lay a foundation for further studies to develop bioinformatic algorithms or to improve the detection of different biologically relevant post-translational modifications.

Acknowledgments This study was supported by grants from the Department of Research and Development at the Karolinska University Hospital and Stockholm County Council, Sweden, the Cancer Society in Stockholm, Sweden, Swedish Cancer Society and the Swedish Research Counsel 2009-5083 and 2004-5259. Instrumentation was acquired through grants from Knut and Alice Wallenberg Foundation.

References

Ahrne E, Muller M, Lisacek F (2009) Unrestricted identification of modified proteins using MS/MS. Proteomics 10(4):671–686

Anderson NL, Jackson A, Smith D, Hardie D, Borchers C, Pearson TW (2009) Siscapa peptide enrichment on magnetic beads using an in-line bead trap device. Mol Cell Proteomics 8(5):995–1005

Bellew M, Coram M, Fitzgibbon M, Igra M, Randolph T, Wang P, May D, Eng J, Fang R, Lin C, Chen J, Goodlett D, Whiteaker J, Paulovich A, McIntosh M (2006) A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC–MS. Bioinformatics 22(15):1902–1909

Berg M, Parbel A, Pettersen H, Fenyo D, Bjorkesten L (2006) Detection of artifacts and peptide modifications in liquid chromatography/mass spectrometry data using two-dimensional signal intensity map data visualization. Rapid Commun Mass Spectrom 20(10):1558–1562

Blondelle SE, Ostresh JM, Houghten RA, Perez-Paya E (1995) Induced conformational states of amphipathic peptides in aqueous/lipid environments. Biophys J 68(1):351–359

Browne CA, Bennett HP, Solomon S (1982) The isolation of peptides by high-performance liquid chromatography using predicted elution positions. Anal Biochem 124(1):201–208

Cargile BJ, Stephenson JL Jr (2004) An alternative to tandem mass spectrometry: isoelectric point and accurate mass for the identification of peptides. Anal Chem 76(2):267–275

Cargile BJ, Bundy JL, Freeman TW, Stephenson JL Jr (2004a) Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. J Proteome Res 3(1):112–119

Cargile BJ, Talley DL, Stephenson JL Jr (2004b) Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pi predictability of peptides. Electrophoresis 25(6):936–945

Cargile BJ, Sevinsky JR, Essader AS, Stephenson JL Jr, Bundy JL (2005) Immobilized pH gradient isoelectric focusing as a first-dimension separation in shotgun proteomics. J Biomol Tech 16(3):181–189

Cargile BJ, Sevinsky JR, Essader AS, Eu JP, Stephenson JL Jr (2008) Calculation of the isoelectric point of tryptic peptides in the pH 3.5–4.5 range based on adjacent amino acid effects. Electrophoresis 29(13):2768–2778

Craig R, Cortens JP, Beavis RC (2004) Open source system for analyzing, validating, and storing protein identification data. J Proteome Res 3(6):1234–1242

Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, Fausto N, Hafen E, Hood L, Katze MG, Kennedy KA, Kregenow F, Lee H, Lin B, Martin D, Ranish JA, Rawlings DJ, Samelson LE, Shiio Y, Watts JD, Wollscheid B, Wright ME, Yan W, Yang L, Yi EC, Zhang H, Aebersold R (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol 6(1):R9. doi:gb-2004-6-1-r9[pii]10.1186/gb-2004-6-1-r9

Eriksson H, Lengqvist J, Hedlund J, Uhlen K, Orre LM, Bjellqvist B, Persson B, Lehtio J, Jakobsson PJ (2008) Quantitative membrane proteomics applying narrow range peptide isoelectric focusing for studies of small cell lung cancer resistance mechanisms. Proteomics 8(15):3008–3018

Essader AS, Cargile BJ, Bundy JL, Stephenson JL Jr (2005) A comparison of immobilized pH gradient isoelectric focusing and strong-cation-exchange chromatography as a first dimension in shotgun proteomics. Proteomics 5(1):24–34

Fenyö D, Beavis RC (2008) Informatics development: challenges and solutions for MALDI mass spectrometry. Mass Spectrom Rev 27(1):1–19

Fraterman S, Zeiger U, Khurana TS, Rubinstein NA, Wilm M (2007) Combination of peptide OFFGEL fractionation and label-free quantitation facilitated proteomics profiling of extraocular muscle. Proteomics 7(18):3404–3416

Gauci S, van Breukelen B, Lemeer SM, Krijgsveld J, Heck AJ (2008) A versatile peptide p*I* calculator for phosphorylated and N-terminal acetylated peptides experimentally tested using peptide isoelectric focusing. Proteomics 8(23–24):4898–4906

Gilar M, Jaworski A, Olivova P, Gebler JC (2007) Peptide retention prediction applied to proteomic data analysis. Rapid Commun Mass Spectrom 21(17):2813–2821

Gupta N, Tanner S, Jaitly N, Adkins JN, Lipton M, Edwards R, Romine M, Osterman A, Bafna V, Smith RD, Pevzner PA (2007) Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. Genome Res 17(9):1362–1377

Heller M, Michel PE, Morier P, Crettaz D, Wenz C, Tissot JD, Reymond F, Rossier JS (2005a) Two-stage Off-Gel isoelectric focusing: protein followed by peptide fractionation and application to proteome analysis of human plasma. Electrophoresis 26(6):1174–1188

Heller M, Ye M, Michel PE, Morier P, Stalder D, Junger MA, Aebersold R, Reymond F, Rossier JS (2005b) Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides. J Proteome Res 4(6):2273–2282

Horth P, Miller CA, Preckel T, Wenz C (2006) Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis 10.1074/mcp.T600037-mcp200. Mol Cell Proteomics 5(10):1968–1974

Kay R, Barton C, Ratcliffe L, Matharoo-Ball B, Brown P, Roberts J, Teale P, Creaser C (2008) Enrichment of low molecular weight serum proteins using acetonitrile precipitation for mass spectrometry based proteomic analysis. Rapid Commun Mass Spectrom 22(20):3255–3260

Kino K, Sugiyama H (2001) Possible cause of g–c→c–g transversion mutation by guanine oxidation product, imidazolone. Chem Biol 8(4):369–378

Krijgsveld J, Gauci S, Dormeyer W, Heck AJ (2006) In-gel isoelectric focusing of peptides as a tool for improved protein identification. J Proteome Res 5(7):1721–1730

Krokhin OV (2006) Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-a pore size c18 sorbents. Anal Chem 78(22):7785–7795

Krokhin OV, Craig R, Spicer V, Ens W, Standing KG, Beavis RC, Wilkins JA (2004) 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 3(9):908–919

Lam H, Aebersold R (2010) Spectral library searching for peptide identification via tandem MS. Methods Mol Biol 604:95–103

Lengqvist J, Uhlen K, Lehtio J (2007) iTRAQ compatibility of peptide immobilized pH gradient isoelectric focusing. Proteomics 7(11):1746–1752

Mant CT, Burke TW, Black JA, Hodges RS (1988) Effect of peptide chain length on peptide retention behaviour in reversed-phase chromatography. J Chromatogr 458:193–205

Matthiesen R, Trelle MB, Hojrup P, Bunkenborg J, Jensen ON (2005) Vems 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. J Proteome Res 4(6):2338–2347

Meek JL (1980) Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. Proc Natl Acad Sci USA 77(3):1632–1636

Nielsen ML, Savitski MM, Zubarev RA (2006) Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. Mol Cell Proteomics 5(12):2384–2391

Norbeck AD, Monroe ME, Adkins JN, Anderson KK, Daly DS, Smith RD (2005) The utility of accurate mass and LC elution time information in the analysis of complex proteomes. J Am Soc Mass Spectrom 16(8):1239–1249

Olsen JV, Mann M (2004) Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. Proc Natl Acad Sci USA 101(37):13417–13422

Savitski MM, Nielsen ML, Zubarev RA (2006) Modificomb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. Mol Cell Proteomics 5(5):935–948

Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA (2007) The paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra 10.1074/mcp.T600050-mcp200. Mol Cell Proteomics 6(9):1638–1655

Spicer V, Yamchuk A, Cortens J, Sousa S, Ens W, Standing KG, Wilkins JA, Krokhin OV (2007) Sequence-specific retention calculator. A family of peptide retention time prediction algorithms in reversed-phase HPLC: applicability to various chromatographic conditions and columns. Anal Chem 79(22): 8762–8768

Stephenson JL Jr, Bunger MK, Cargile BJ, Sevinsky JR (2006) A new algorithm for pi prediction of peptides from IPG-IEF: applications to analysis of single nucleotide polymorphisms. In: 7th Siena meeting from genome to proteome: back to the future, Siena, Italy, 3–7 September 2006

Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V (2005) Inspect: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 77(14):4626–4639

Uhlén K, Fenyo D, Hörnsten L, Bjellqvist B (2006) Improved prediction of peptide isoelectric point by modelling the effects of interaction between charged neighbouring amino acids. In: 7th Siena meeting from genome to proteome: back to the future, Siena, Italy, 3–7 September 2006

Vaezzadeh AR, Hernandez C, Vadas O, Deshusses JJ, Lescuyer P, Lisacek F, Hochstrasser DF (2008) PICarver: a software tool and strategy for peptides isoelectric focusing. J Proteome Res 7(10):4336–4345

Wessel D, Flugge UI (1984) A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. Anal Biochem 138(1):141–143

Wisniewski JR, Zougman A, Nagaraj N, Mann M (2009) Universal sample preparation method for proteome analysis. Nat Methods 6(5):359–362

Xie H, Gilar M, Gebler JC (2009) Characterization of protein impurities and site-specific modifications using peptide mapping with liquid chromatography and data independent acquisition mass spectrometry. Anal Chem 81(14):5699–5708

Zubarev R, Mann M (2007) On the proper use of mass accuracy in proteomics. Mol Cell Proteomics 6(3):377–381

Zybailov B, Sun Q, van Wijk KJ (2009) Workflow for large scale detection and validation of peptide modifications by RPLC-LTQ-Orbitrap: application to the *Arabidopsis thaliana* leaf proteome and an online modified peptide library. Anal Chem 81(19):8015–8024