research articles Journal of proteome research

In-Gel Isoelectric Focusing of Peptides as a Tool for Improved Protein Identification

Jeroen Krijgsveld,* Sharon Gauci, Wilma Dormeyer, and Albert J. R. Heck

Department of Biomolecular Mass Spectrometry, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

Received March 27, 2006

In the analysis of proteins in complex samples, pre-fractionation is imperative to obtain the necessary depth in the number of reliable protein identifications by mass spectrometry. Here we explore isoelectric focusing of peptides (peptide IEF) as an effective fractionation step that at the same time provides the added possibility to eliminate spurious peptide identifications by filtering for p*I*. Peptide IEF in IPG strips is fast and sharply confines peptides to their p*I*. We have evaluated systematically the contribution of p*I* filtering and accurate mass measurements on the total number of protein identifications in a complex protein mixture (*Drosophila* nuclear extract). At the same time, by varying Mascot identification cutoff scores, we have monitored the false positive rate among these identifications by searching reverse protein databases. From mass spectrometric analyses at low mass accuracy using an LTQ ion trap, false positive rates can be minimized by filtering of peptides not focusing at their expected p*I*. Analyses using an LTQ-FT mass spectrometer delivers low false positive rates by itself due to the high mass accuracy. In a direct comparison of peptide IEF with SDS-PAGE as a pre-fractionation step, IEF delivered 25% and 43% more proteins when identified using FT−MS and LTQ−MS, respectively. Cumulatively, 2190 non redundant proteins were identified in the *Drosophila* nuclear extract at a false positive rate of 0.5%. Of these, 1751 proteins (80%) were identified after peptide IEF and FT-MS alone. Overall, we show that peptide IEF allows to increase the confidence level of protein identifications, and is more sensitive than SDS-PAGE.

Keywords: false positive rate • isoelectric focusing • SDS-PAGE • FT−MS • LTQ−MS • *Drosophila*

Introduction

The success of large-scale proteomics studies is strongly dependent on the ability to identify large numbers of proteins in biologically complex mixtures. Recent developments in mass spectrometry, especially in conjunction with liquid chromatography, have greatly enhanced the capabilities to characterize such samples in considerable depth. The online coupling of capillary reversed phase chromatography to electrospray mass spectrometry has particularly emerged as a powerful technique enabling the separation and identification of hundreds of peptides in a single experiment.[1−3]

Despite this progress, the huge complexity and dynamic range of proteins in biological samples in general still largely exceeds the analytical capabilities of the mass spectrometer, even in combination with chromatographic separation at the front.

Although modern mass spectrometers with increased scanning speeds and shorter duty cycles are a distinct improvement, an alternative solution to reduce sample complexity lies in protein or peptide pre-fractionation. A myriad of possibilities is available that, especially when combined in multidimensional systems, provides tools for the "comprehensive" characterization of a sample. Most strategies are orthogonal to the "standard" combination of LC−MS,[4] either at the protein[5−7] or the peptide level.[8] For instance, peptide fractionation by SCX−LC followed by reversed-phase-LC, termed MudPit,[9] is a well-established approach for peptide separation enabling the identification of multiple thousands of peptides in a single experiment. An alternative and common way to separate proteins is SDS polyacrylamide gel electrophoresis (PAGE). Here, proteins are separated by size, which can be excised for in-gel proteolytic digestion.[10] Depending on the sample and the question to be addressed, many alternatives are available, including two-dimensional separation of proteins,[7] anion-exchange chromatography but also subcellular fractionation[11] and affinity purification, to name just a few.

Another issue in proteomics lies in the process of identification of proteins in complex mixtures, and specifically in defining confidence scores for unattended identification. Usually, a software algorithm is used to match an experimental peptide fragmentation pattern against a database of all proteins potentially present in a sample.[12] In large data sets, multiple thousands of spectra can be matched against the entire

* To whom correspondence should be addressed. Utrecht University, Department Biomolecular Mass Spectrometry, Sorbonnelaan 16, 3584 CA Utrecht, The Netherlands. Tel: +31-30-2539974. Fax: +31-30-2518219. E-mail: j.krijgsveld@chem.uu.nl.

10.1021/pr0601180 CCC: $33.50

Published on Web 06/09/2006

proteome of the organism under investigation. For each experimental spectrum the best matching peptide is scored reflecting the quality of the match. Usually, a threshold value is used to discriminate correct from false identifications. The problem is that the various search engines (e.g., Sequest, Mascot, Spectrum Mill, ProID) use their own scoring algorithm and can hardly be compared.

This may be illustrated by the fact that submission of the same data set to various search engines does not give identical answers,[13,14] leaving a degree of uncertainty on the correctness of the identifications. It has therefore been suggested that only identifications should be reported that were identified by 2 or more algorithms,[14] or to validate identifications manually[15] or, more realistically, by statistical models.[16,17] What may be learned from these and other studies is that even in abundantly used algorithms no consensus seems to exist as to what settings are to be used to obtain valid identifications at the desired sensitivity and specificity levels.

Assignment of fragmentation spectra to peptide sequences can be solely based on scoring values, but various other orthogonal criteria have been explored to accept or refute identifications. One of these is the retention time of peptide during reversed phase chromatography. Within certain experimental settings, chromatographic behavior of peptides can be predicted, to be used as an additional criterion for their identification.[18] Another parameter explored more recently is the isoelectric point of peptides, which can be exploited after isoelectric focusing (IEF) of peptides. Recently, this has been investigated for peptides in solution,[19] by free flow electrophoresis (FFE),[20] by capillary electrophoresis[21] and by in-gel electrophoresis in IPG-strips.[22] Each of these approaches benefit from the general advantage of IEF that peptides can be binned and concentrated in discrete fractions, which is generally not possible by chromatographic techniques. However, their applicability in practice also depends on the size of fractions that can be sampled, presence of carrier ampholytes that could complicate further analysis, and minimum sample loads that are required. Stephenson and co-workers[23,24] and others[20] have shown that the expected p*I* of a peptide can be used to support or refute its identification as determined by mass spectrometry.

In this paper, we have addressed the question to what extent filtering on p*I* contributes to the quality of large data sets. We have taken a *Drosophila* nuclear extract as a model for a complex protein mixture, and have analyzed this by peptide IPG-IEF followed by LC−MS. To get a sense how this compares to other frequently used techniques, we have run a parallel experiment where the same sample was analyzed by protein SDS-PAGE followed by in gel digestion and LC−MS. Our particular interest was whether peptide IEF performed at the desired sensitivity level to be a viable alternative for protein SDS-PAGE. We have addressed this also in the light of accurate mass measurements, another aspect aiding in confident protein identification. We have used reverse sequence protein databases as a validation tool to estimate the false positive rate in the various data sets. In this analysis, we show that data filtering based on pI is a highly valuable tool to increase the confidence in peptide and protein identifications, especially for data generated on instrumentation delivering lower mass accuracy. Besides the increased confidence also the proteome coverage could be improved since compared to protein SDS-PAGE, peptide IEF delivered 25−43% more protein identifications.

Materials and Methods

Peptide IEF. Drosophila nuclear extract was a gift from Dr. C. P. Verrijzer (ErasmusMC, Rotterdam, The Netherlands) and was prepared according to Heberlein and Tjian.[25] The final preparation was dissolved in HEMG (25 mM HEPES−KOH pH 7.6, 0.1 mM EDTA, 12.5 mM $MgCl_2$ and 10% glycerol) at a protein concentration of 5 mg/mL. A 100-μg portion of protein was diluted with 50 mM ammonium bicarbonate (pH 8) in a total volumne of 50 μL and was digested with trypsin in a enzyme:substrate ratio of 1:40 (w/w) at 37 °C for 16 h. The sample was brought up to 8 M urea in the presence of IPG buffer 3−10 NL (Amersham) and applied to a 13 cm IPG dry strip, 3−10 NL (Amersham). Using an IPGphor (Amersham), the following focusing protocol was applied: 16 h 30 V, 3 h 500 V, 8000 V up to 30 000 Vh. Alternative settings used during the optimization of IEF are indicated in the text. Excess cover oil was removed, and the gel was scraped off the plastic backing in 20 equally sized parts within 3 min to prevent diffusion. Peptides were eluted in three sequential solutions containing 0%, 50%, and 100% acetonitril in water and 0.1% TFA. Supernatants were combined, dried down and redissolved in 5% acetic acid. Samples were cleaned of salts and residual oil using STAGE tips,[26] dried, and stored at −80 °C before analysis. A script written in-house was used to calculate the pI of peptide batch wise, employing an algorithm that is also used at Expasy (http://www.expasy.org/tools/pi_tool.html).

Mass Spectrometry. Nanoflow-LC tandem mass spectrometry was performed by coupling an Agilent 1100 HPLC (Agilent Technologies), operated as described before,[27] to a 7-tesla LTQ-FT mass spectrometer (Thermo Electron, Bremen, Germany) or an LTQ ion trap (Thermo Electron, Bremen, Germany). For peptide LC, trapping columns (1 cm × 100 μm) and analytical columns (15 cm × 50 μm) were packed in-house with ReproSil-Pur C18-AQ, 3 μm (Dr. Maisch GmbH, Ammerbuch, Germany). Peptide mixtures were delivered at 3 μL/min on the trapping column for desalting. After flow-splitting down to ~150 nL/min, peptides were transferred to the analytical column and eluted in a gradient of acetonitrile (1%/min) in 0.1 M acetic acid. The eluent was sprayed via emitter tips (New Objective), butt-connected to the analytical column.

Mass spectrometers were operated in data dependent mode, automatically switching between MS and MS/MS acquisition for the three most abundant peaks in a given MS spectrum. In the LTQ-FT, full scan MS spectra were acquired in the FT−ICR at a target value of 5E6 with a resolution of 20 000. The three most intense ions were then isolated for accurate mass measurements by a FT−ICR selected ion monitoring scan which consisted of 10 Da mass range, at a resolution of 50 000. These ions were then fragmented in the linear ion trap. In the LTQ, MS scans were recorded in centroid mode at a target value of 30 000. Peptides were fragmented when the signal exceeded 2E4 counts by filling the ion trap at a target value of 10 000 with a maximum ion time of 100 ms. From MS/MS data in each LC run peak lists were created using Bioworks 3.1 software (Thermo Electron, Bremen, Germany) which were converted to a single file in Mascot generic format using a perl script written in-house.

Database Searching. For protein identification, database searches were performed using Mascot version 2.0, allowing 5 ppm and 1.2 Da mass deviation for the precursor ion for data generated by the LTQ-FT and the LTQ, respectively. Methionine oxidation and cysteine carbaminomethylation were allowed as variable and fixed modifications, respectively. The *Drosophila*

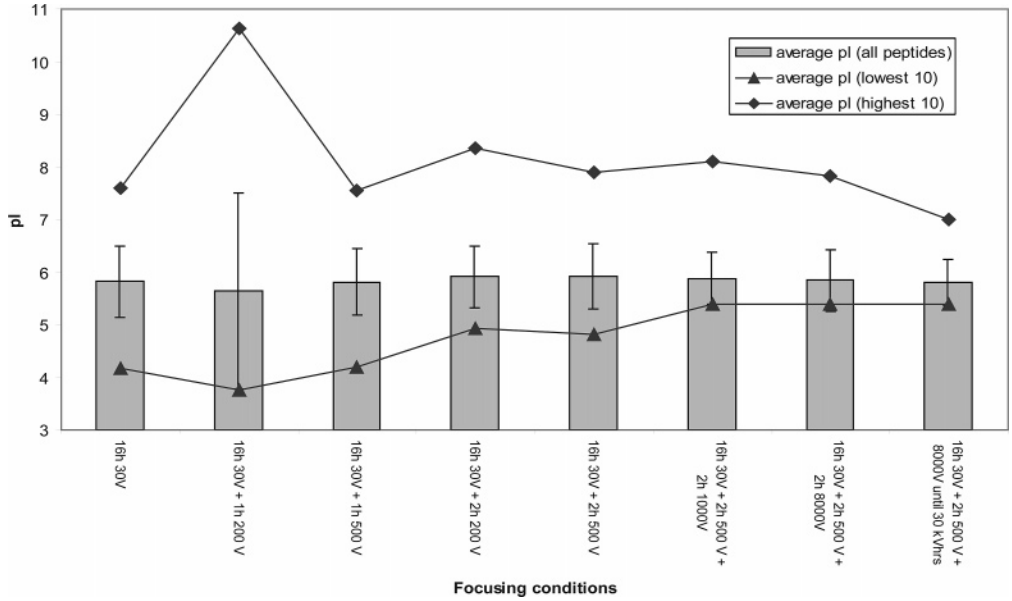

Figure 1. Optimization of peptide isoelectric focusing. Average theoretical pI values for peptides identified after IPG using focusing conditions as indicated ±SD (grey bars). Average p*I* of the 10 peptides with the highest (squares) and lowest p*I* (triangles) are indicators for the occurrence of outliers. Results are shown for one representative fraction (fraction 9 out of 20).

proteome database (ftp://ftp.ebi.ac.uk/pub/databases/integr8/fasta/proteomes/) was used for protein identifications. Mascot cutoff scores were applied as indicated in the article. To estimate the number of false positive identifications under the various experimental conditions we made use of reversed databases throughout the analysis. A reversed database contains all proteins a 'true' database, but with all sequences inversed. False positive estimations rely on the assumption that any protein identification from a reversed database is false.[28,29] In our analyses, a composite database was used with all protein sequences of the Drosophila proteome database in forward as well as in reverse direction. False positive rates were calculated using the eq 2 * n(rev)/n(rev) + n(forw), where n(forw) and n(rev) are the number of peptides identified in proteins with forward (normal) and reversed sequence, respectively.

Results

Optimization of Experimental Conditions for IEF. Since optimal conditions for peptide IEF in IPG strips have not been reported yet, IEF settings were initially taken directly from 2D-gel applications for proteins (i.e., 30 kVh at 8000 V). To gain further insight into the required running conditions for in-gel peptide IEF, various time and voltage courses were tested. Similar aliquots of 50 μg of trypsin-digested nuclear extracts of *Drosophila* embryos were applied to eight 13 cm IPG dry strips, 3−10 NL, and focusing was performed as follows using 8 different conditions: 1: 16 h 30 V (rehydration only); 2: 16 h 30 V, 1 h 200 V; 3: 16 h 30 V, 1 h 500 V; 4: 16 h 30 V, 2 h 200 V; 5: 16 h 30 V, 2 h 500 V; 6: 16 h 30 V, 2 h 500 V, 2 h 1000 V; 7: 16 h 30 V, 2 h 500 V, 2 h 8000 V; 8: 16 h 30 V, 2 h 500 V, 8000 V up to 30 kVh. The runs were conducted independently from each other to prevent mutual influences of the focusing experiments. From each strip 20 slices of gel were scraped off of the plastic backing (slice 1 = acidic end pH 3, slice 20 = basic end pH 10). The peptides were eluted, desalted, concentrated, and analyzed by nanoflow-LC coupled to LTQ ion trap MS. The pI of unique peptides belonging to proteins identified by Mascot database searches were calculated. Optimal focusing conditions were considered to be reflected in a low standard deviation of the average p*I* and in a narrow interval between the most extreme p*I*s of the peptides found in one gel slice. For the latter value, the average of the 10 highest p*I*s and the 10 lowest p*I*s in each of the gel slices was calculated in order to minimize the effect of outliers with extreme p*I*s. This is exemplified for gel slice number 9 as a typical representation of the 20 slices. Across all 8 conditions, 272 ± 47 unique peptides were identified in this section. Surprisingly, decent peptide focusing took already place after rehydration only (30V for 16 h) (Figure 1) with an average p*I* of 5.82 ± 0.68. Extended focusing for 1 h at 200 V or 500 V slightly deteriorated focusing efficiency, while prolonged focusing at higher voltages gradually resulted in a decreased standard deviation of p*I* values. Furthermore, the curves for the average p*I* of the highest 10 and the lowest 10 p*I* converged when higher voltages were applied, indicating the elimination of (extreme) outliers (Figure 1). Proper focusing, i.e., standard deviation between 0.49 and 0.65, seems to be achieved more easily in the acidic half compared to the basic half of the strip. Application of 1000 V or 8000 V was sufficient to obtain optimal standard deviation at low p*I*, whereas at the basic end (slices 16−20) only a standard deviation of 1.6−1.8 was achieved, even after 30 kVh focusing. This already indicated the exceptional behavior of peptides with extreme pI in the in-gel IEF of peptides (see below).

Peptide Separation by Isoelectric Focusing and Analysis by FT−MS. Next, we have investigated the utility of peptide separation based on pI for the identification of proteins in complex mixtures in detail. A nuclear protein isolate (100 μg) of *Drosophila* embryos was digested overnight with 2.5 μg of trypsin, and the peptide mixture was brought up to 8 M urea in the presence of 0.5% IPG buffer. After IEF (30 kVh), the gel was divided into 20 equal parts. Of the extracted peptides in each fraction, 10% was injected for analysis on an LC−LTQ−FT−MS system. The total data set of these 20 LC−MS/MS runs was subjected to a thorough analysis, consisting of the following: (1) the identification of peptides in each fraction, (2) the

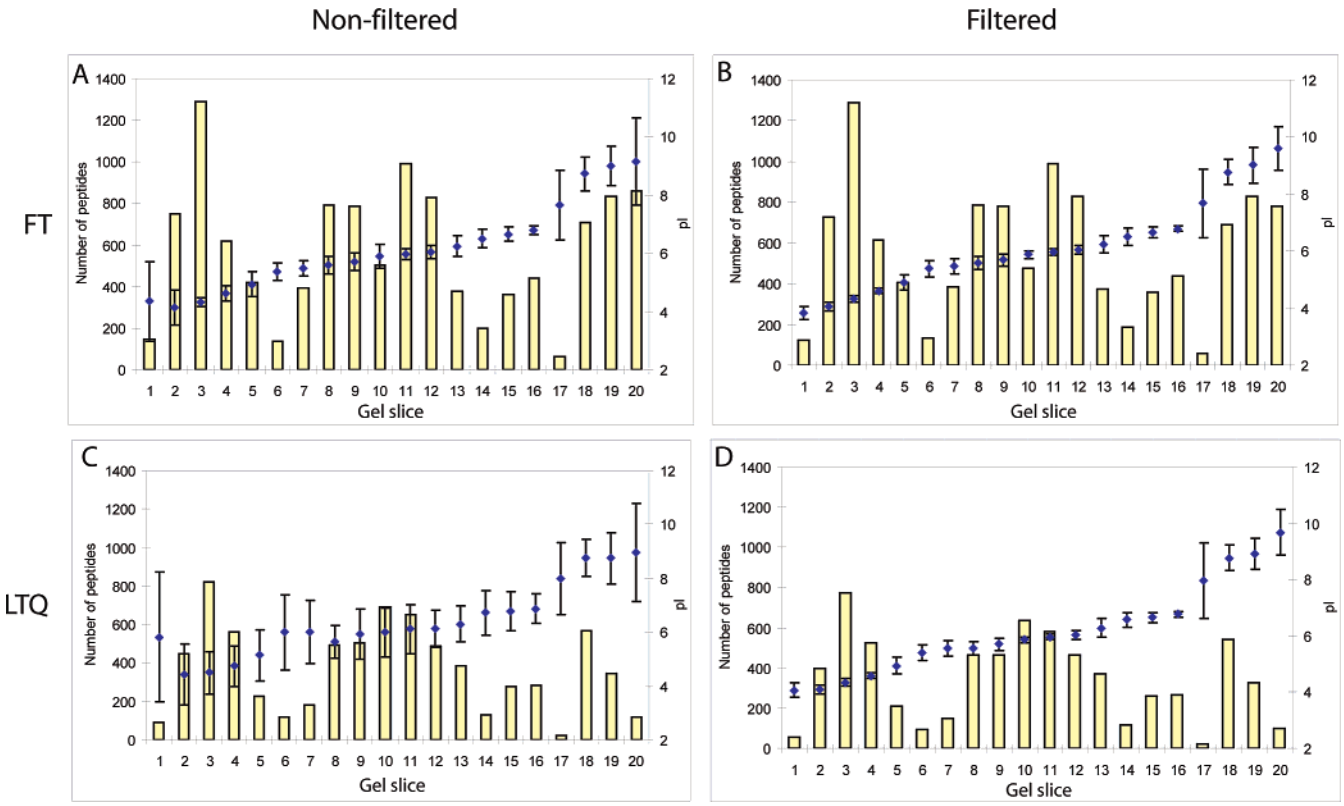

Figure 2. Distribution of tryptic peptides from a *Drosophila* nuclear extract over a 3−10 NL IPG strip (p*I* 3 to 10 in a nonlinear gradient). For each of the 20 gel fractions the number of identified peptides (bars, left axis) and the average p*I* ± SD of these peptides (blue squares, right axis) are plotted. This was done after FT−MS (panels A and B) and after LTQ−MS (panels C and D), both before filtering (panels A and C) and after filtering for p*I* (panels B and D).

Table 1. Comparison of the Number of Peptides Identified by Reversed-Phase Nano-LC by FT−MS and LTQ−MS after Peptide−IEF and Protein SDS-PAGE Gel Electrophoresis as a First Separation Step[a]

	IEF−FT	IEF-LTQ	SDS−FT	SDS-LTQ
before filtering				
#forw peptides	11462	7388	9396	6683
#unique forw pep	9265	6323	7271	5460
#rev pept	159	385	47	127
#unique rev pep	127	375	47	127
%false positives	2.7	9.9	1.	3.7
peptide length forw	13.9	13.2	12.4	13.6
peptide length rev	7.8	9.8	7.6	9.8
after filtering				
#forw peptides	11258	6797		
#unique forw pep	9212	5908		
#rev pept	115	104		
#unique rev pep	106	99		
%false positives	2.0	3.0		

[a] Numbers are shown before and after p*I* filtering (IEF only). Forw pep: peptides with forward sequence, identified from normal database. Rev pep: peptides with reverse sequence, identified from database with all sequences inverted.

distribution of peptides along the IPG gel in relation to their theoretical p*I*, (3) the efficacy of focusing by monitoring the occurrence of peptides in multiple sections of the gel, (4) the ability to use IEF as a criterion to eliminate misidentifications, (5) the number of false positive identifications before and after p*I* filtering.

Peptides in each of the 20 IEF fractions identified by FT−MS data sets are listed in Supporting Information Table 1A, along with their theoretical p*I*. In Figure 2, where these data are plotted in a more concise form, it is shown that the number of peptides ranges from 60 in fraction 17 to 1291 in fraction 3, which readily illustrates that the distribution over the entire pH range is not even. This phenomenon has been observed before for samples of different origin, with remarkably few data points between pH 7 and 8.[22] Collectively, in these 20 fractions 11 621 peptides were identified (reverse sequences included).

Figure 2A also shows that in most of the fractions the p*I* of the identified peptides tightly clusters around p*I* values that gradually increase along the strip. In fact, this matches the expected trend in the nonlinear pH gradient used in this experiment, with steep gradients at the extremes and a more shallow gradient between pH 5 and 7.

To look into p*I* distribution per fraction in more detail, theoretical p*I* values for all peptides are plotted against Mascot identification scores in Supporting Information Table 1A. In Figure 3A the results are shown for a representative fraction (fraction 10). In this fraction, 500 peptides were identified with an average p*I* of 5.89 ± 0.2. In this particular fraction, 485 forward and 15 reverse peptides were found, so that the false positive (FP) rate can be estimated at 2*15/(15 + 485) = 6%. For the entire data set of all 20 fractions the FP rate is 2*159/(159 + 11462) = 2.7%.

We next addressed the question what the effect on the overall FP rate would be when filtering out peptides with deviating p*I*. To this end, we set a pI window for each fraction (Supporting Information Table 2), and removed all peptides outside this window. In Figure 3A,B (fraction 10) and Supporting Information Figure 1 (all fractions) unfiltered and filtered data are given side by side, showing that both false identifications (pink) and

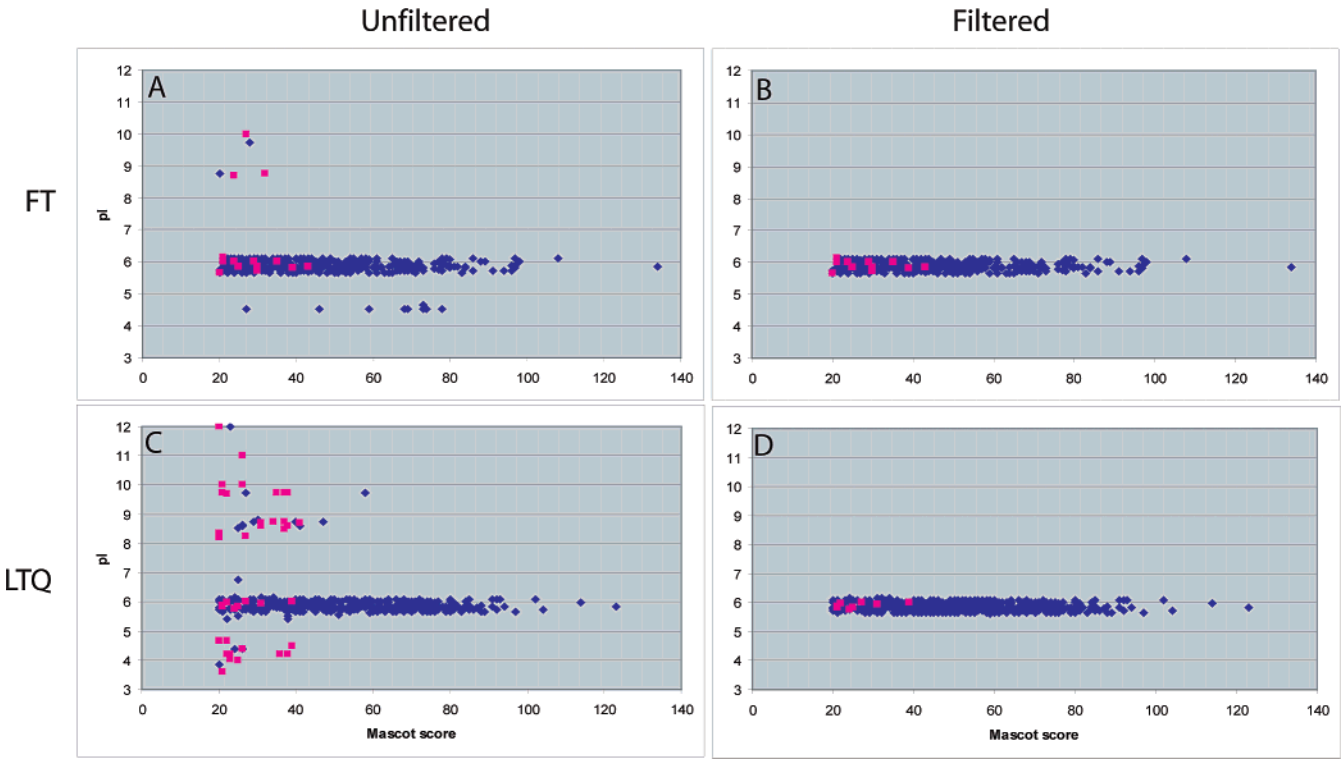

Figure 3. p*I* distribution of peptides identified in fraction 10 of the IEF gel after FT−MS (panels A and B) and LTQ−MS (panels C and D), before filtering (panels A and C) and after filtering for p*I* (panels B and D).

initially presumed correct identifications (blue) are removed in the filtering process. In doing so, the total number of identified peptides went down from 11 621 to 11 372 (including 115 reverse peptides), but, more importantly, the false positive rate was reduced from 2.7 to 2.0%. The benefit for individual fractions is shown in Figure 2, where the average standard deviation in p*I* decreased from 0.49 to 0.31. A remark should be made about fractions 17 to 20 where no filtering was performed since the number of peptides was too small to properly determine filtering conditions (fraction 17), or since the scatter in theoretical p*I* was too high.[18−20] The latter could be due to poor focusing conditions near the extreme end of the gel. A similar phenomenon, but to a lesser extent, can be observed for fraction 1.

Peptide Separation by Isoelectric Focusing and Analysis by LTQ−MS. Although this procedure shows that the confidence of peptide identification can be improved by filtering of peptides based on p*I*, the overall improvement was rather modest. In fact, the false positive rate was already very low in the first place (2.7%), presumably owing to the high mass accuracy that was realized using the FT mass spectrometer. The contribution of p*I* filtering can be expected to be much greater at lower mass accuracies. Therefore, we used the same 20 samples and analyzed these using an iontrap LTQ instead of an FT instrument. Conditions for nanoflow chromatography were identical as above, the main difference was that precursor masses were determined in the LTQ mass spectrometer with an accuracy of ~1.2 Da. With all identified peptides listed in Supporting Information Table 1B, Figure 2C shows that the number of identified peptides in each of the fractions is somewhat lower compared to the FT data (Figure 2A). In addition, the number of false positives in most of the fractions is higher (Supporting Information Figure 2), resulting in higher standard deviations (Figure 2C) (average SD 1.01 pH unit). The entire data set of 7773 peptides, including 385 from the reversed database, contained 9.9% false positives. Removal of p*I*-outliers, using the same filtering conditions as used in the previous analysis (Supporting Information Table 2) resulted in the elimination of 872 peptides, a decrease of the average SD (to 0.31 pH units) and a decrease of the FP rate to 3.0%.

From these data, it appears that analysis of the same sample using two different instruments (FT and LTQ) delivers datasets that differ considerably both in the number of peptides (11621 vs 7773) and FP rate (2.7 vs 9.9%). Importantly, the latter numbers drop to fully acceptable levels for both data sets after filtering for p*I* (2.0% and 3.0% for FT and LTQ, resp, Table 1).

Resolving Power of IEF. To make peptide separation by IEF followed p*I* filtering a valuable addition to existing techniques, it is important that focusing occurs in a tight region in the gel. For the peptides identified by FT−MS remaining after filtering we analyzed the frequency of each peptide across the 20 gel fractions. In the total set of 11 258 peptides, 7679 peptides were found only once (Figure 4, note the log scale), 1290 were found in 2 fractions, 296 in 3 fractions, and decreasing numbers were identified in up to 6 fractions. This means that 96% of all peptides were found in two fractions at maximum (82% in a single fraction, 14% in two fractions). Another important observation is that all peptides identified more than once were always found in adjacent fractions. Inspection of the peptides that were found more than 3 times learned that they represent some high abundant proteins, such as transcription elongation factors, heat shock proteins, ribosomal proteins and vitellogenins (yolk proteins). Sub-optimal focusing may be due to peptide abundance and local overloading of the strip.

Comparison of Peptide-IEF with Protein SDS-PAGE as a Separation Step Prior to LC−MS. An alternative routine approach to identify proteins in complex mixtures is the separation of proteins by SDS-PAGE followed by in-gel diges-

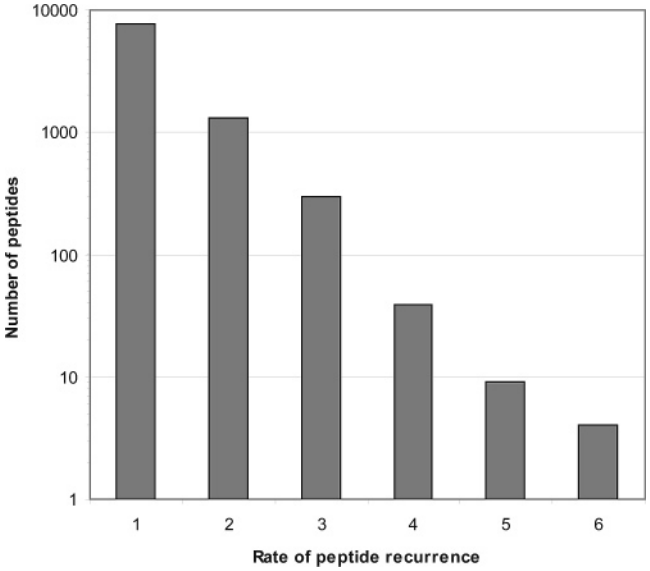

Figure 4. Peptide redundancy between IEF fractions. *x*-axis: frequency of a particular peptide identification across 20 IEF-fractions. A log scale is used to indicate the sharp decline in the number of redundant identifications.

tion of excised parts and LC−MS/MS analysis.[10] We have compared peptide IEF with protein SDS-PAGE as a separation step prior to LC−MS. To make it fully comparable, we have analyzed the same amount of protein, digested the gel in the same number of slices, injected the same amount of sample (10% of each digest) into the LC system and used the same methods and settings for MS, both for the FT and the LTQ. Peptides identified in the 20 SDS-gel fractions by FT−MS and LTQ−MS are listed in Supporting Information Table 1C,D, respectively. From the distribution of the number of identifications over the gel (Supporting Information Figure 3) it appears that most peptides are found in the high-molecular weight fractions, with peptide numbers ranging from 28 (fraction 17) to over 800 (fraction 2). In total, 9396 and 6683 peptides were identified by FT−MS and LTQ−MS, resp (Table 1). These numbers are 22% and 10% lower compared to IEF as a first separation step. When considering only unique peptides in these sets, IEF outperforms SDS-PAGE by 27% and 16% using FT−MS and LTQ−MS, respectively (Table 1). After SDS-PAGE, FP rates are relatively low (1.0% and 3.7%, for FT and LTQ).

Peptide IEF and False Positive Rates in Protein Identifications. So far, we have only considered numbers of identifications and false positive rates at the peptide level. This is only the first step in protein identification, and thus it is far more interesting to consider these entities at the protein level. Our aim was to generate the best possible list of identifications in the nuclear extract containing the largest number of proteins as possible with the lowest possible false positive rate. As a first step to generate such a list we removed redundant peptides, and excluded peptides with a Mascot score below 25. Then, peptides belonging to the same protein were combined, and protein scores were calculated as the sum of the peptide scores. Finally, for every data set a protein cutoff score was selected such that a 0.5% FP rate was achieved.

We investigated the quality of the set of identified proteins in the IEF data set before and after p*I*-filtering. In Figure 5A the number of identified proteins and the FP rates are shown as a function of the Mascot protein score before and after p*I* filtering. By increasing the cutoff score from 25 to 60, the number of proteins identified by FT−MS gradually decreases from ∼2200 to ∼1700. Over this interval, the FP rate drops from 6% to virtually 0%. After p*I* filtering, both the number of identifications and FP rates are marginally affected, as was observed for peptides before (Figure 2). Across the cutoff range, the number of proteins identified by LTQ-MS decreased from ∼2000 to ∼1200 (Figure 5A). Filtering for p*I* had little effect on this number, but, importantly, FP rates dropped dramatically from 16% to 6% at a cutoff of 25, and from 1.5% to 0.4% at a cutoff of 60.

A similar trend is observed for proteins identified after SDS electrophoresis (Figure 5B) with higher false positive rates and lower number of identifications after LTQ−MS compared to FT−MS. Since no p*I* filtering can be applied in this case, higher cutoff values are required to obtain a similar confidence level. For instance, a 1% FP rate for proteins identified by LTQ−MS after IEF and p*I* filtering is achieved at a Mascot score of 50 (Figure 5B), whereas after SDS a minimum score of 56 would be required.

To compare the four datasets (peptide IEF and protein SDS-PAGE, each analyzed by LTQ−MS and FT−MS) in a meaningful way we set a desired FP rate to evaluate the number of protein identifications. We chose to use a stringent FP rate of 0.5% to achieve a dataset with high-confident identifications.

Protein cutoff scores needed to realize FP rates of 0.5% in each of the four datasets can be deduced from Figure 5A. These values were 35 (SDS-PAGE followed by FT−MS), 43 (IEF followed by FT−MS), 52 (IEF followed by LTQ−MS) and 59 (SDS-PAGE followed by LTQ−MS) (Figure 6). The variation between these values illustrate that there is no standard cutoff value that can be used if one aims to identify the maximum number of proteins at a given confidence level.

All proteins identified using these criteria in each of the four approaches are listed in Supportng Information Table 3, and the peptides that were identified for these proteins are listed in Supporting Information Table 4. The total number of identifications are summarized in Figure 6. Cumulatively, 2190 nonredundant proteins were identified by 15151 peptides. It is striking to note that in this high-quality data set numbers are clearly in favor of peptide IEF when compared to SDS electrophoresis: 25% (1752/1405) more proteins were identified by FT−MS, and even 43% (1239/864) more by LTQ−MS (Figure 6). The difference between FT and LTQ is also substantial, but may be less unexpected: 41% more proteins are identified after IEF, and 63% more after SDS−PAGE. Of all the proteins, the majority (80%) were identified solely using IEF−FT−MS, reflecting the strength of both IEF and FT−MS. Addition of the 3 other data sets increased the number of proteins, but mainly extends the number of peptides per protein (from 4.6 to 6.9). The added value of IEF and FT−MS can also be seen from the Venn diagrams in Figure 6, where substantially more proteins were identified by these methods compared to SDS-PAGE and LTQ−MS, respectively.

Qualitative Comparison of Peptides and Proteins Identified after IEF and SDS-PAGE. A relevant question that remains is how the four datasets differ at the individual peptide and protein level: which are the ones that are identified after IEF but not SDS electrophoresis? Or are there also examples of proteins solely identified after SDS-PAGE? To get a global overview of the peptide characteristics (obtained from Supporting Information Table 4), the distribution in peptide length and pI in each of the four approaches was investigated (Supporting Information Figure 4). When comparing IEF to

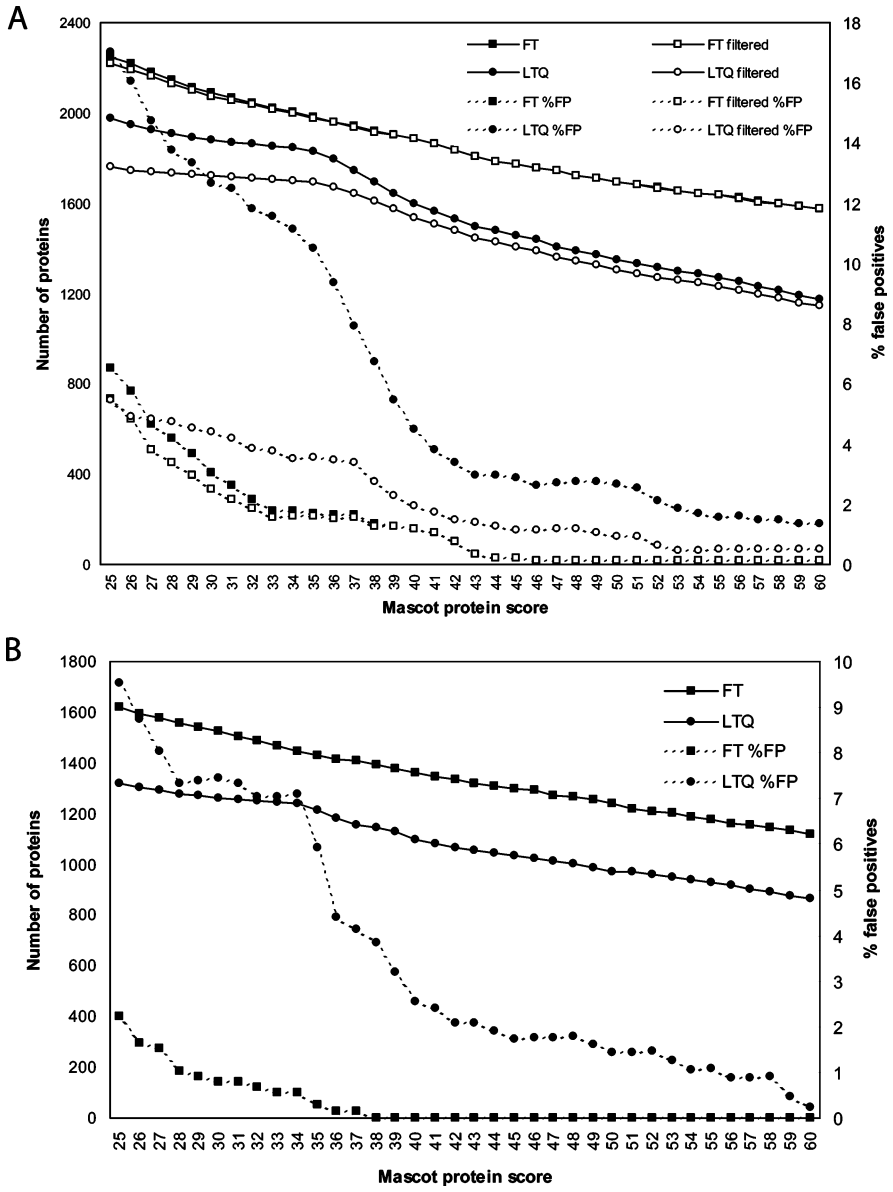

Figure 5. Correlation between mascot protein score, the number of protein identifications and false positive rates. For each protein mascot cutoff score, the number of identified proteins (solid lines, left axis) and the false positive rate (dashed lines, right axis) was determined. Panel A and B show these data for IEF and SDS-PAGE, respectively, followed by FT−MS (squares) and LTQ−MS (circles). Open and solid symbols denote data obtained with and without filtering peptides for p*I*, respectively. Filtering could only be performed after IEF (panel A).

SDS-PAGE as a pre-fractionation step, it appears that in IEF longer peptides (10−20 residues) are clearly overrepresented, while analysis after SDS-PAGE tends to yield more shorter peptides (5−8 residues) (Supporting Information Figure 4A). At the same time, most peptides identified after SDS-PAGE have a p*I* between 4 and 5, whereas the maximum number of peptides after IEF are found between 5 and 7 (Supporting Information Figure 4B). Apparently, the extensive separation in IEF of peptides between pH 5 and 7 (fraction 5−16, see Figure 2) significantly contributes to the total number of identifications.

To get insight as to what (type of) proteins may be preferentially identified after IEF or SDS-PAGE, for each identified protein in Supporting Information Table 3 the cumulative number of peptides after IEF (by FT and LTQ mass spectrometry) and SDS is included. The ratio between these numbers is also provided, to easily trace those proteins that appear preferentially in one of these methods. From this comparison it appears that 711 proteins were identified after IEF, but not SDS-PAGE. On the other hand, the inverse is observed for 298 proteins. This latter observation is at least remarkable, especially taking into consideration that some proteins are identified by as many as 23 peptides (vs no peptides after IEF), indicating that these are present at considerable abundance. Overall, 91 proteins were identified by 3 or more peptides after SDS-PAGE which were completely missed after IEF, while 85 proteins were identified by >3-fold more peptides after IEF compared to SDS-PAGE.

A possible explanation for this discrepancy could reside in differences in sample preparation, causing various peptide yields for proteins differing in physical properties. This may be illustrated by the selective finding of 20S proteasomal proteins (all alpha subunits PSA1 to 7, and beta subunits (PSB) 1,2 and 4) after SDS-PAGE, but not IEF. Interestingly, the

	IEF-FT	IEF-LTQ	SDS-FT	SDS-LTQ	cumul
Mascot cut-off	43	52	35	59	
# unique peptides	8146	5038	6300	4413	15151
# unique proteins	1751	1238	1404	863	2190
peptides/protein	4.65	4.07	4.49	5.11	6.92

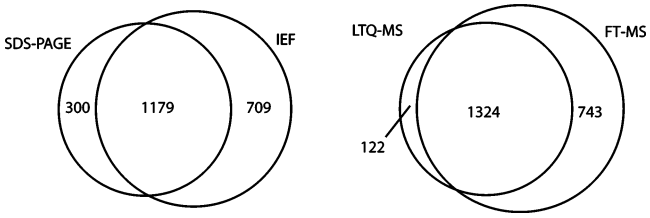

Figure 6. Comparison of the number of proteins identified by reversed-phase nano-LC by FT−MS and LTQ−MS after peptide-IEF and protein SDS-PAGE gel electrophoresis as a first separation step. For IEF only peptides were included retained after p*I* filtering. Cumul denotes the cumulative number of unique peptides and proteins across the four methods. Numbers in the Venn diagrams indicate the overlap in protein identifications after SDS-PAGE and IEF (FT−MS and LTQ−MS accumulated), and after LTQ−MS and FT−MS (IEF and SDS-PAGE accumulated).

subunits from the regulatory 19S domain of the proteasome were primarily found after IEF (RPN 1, 2, 6, 7, 9). This discrepancy could be explained by differences inherent to the various protocols used for sample preparation. Notably, proteins were dissolved and boiled in SDS sample buffer before SDS-PAGE, while proteins were digested under nondenaturing conditions before IEF. Stable protein complexes or highly globular proteins would be refractory to digestion under these conditions, and would thus be missed in subsequent analyses.

Thus, rather than presenting an explanation why more proteins were identified after IEF, a suggestion is provided for improved sample preparation prior to IEF. For instance, proteolysis under denaturing conditions (e.g., in urea) could be a distinct advantage in identifying (even) more proteins using peptide IEF than we have observed in the current analysis.

Discussion

Striving for large numbers of protein identifications at high sensitivity is often in conflict with high MS data quality and confident protein identifications. Nevertheless, it is certainly possible to identify many hundreds of proteins in a single sample using multidimensional separation interfaced with mass spectrometry. Common practice for protein identification is the matching peptide fragmentation patterns against a protein database which can be performed by a variety of search algorithms. Most of these differ in the way they rank protein identifications and in the way they produce probability scores that reflect the quality of each match. What they have in common, however, is that in all cases protein identification is a probabilistic process with the inherent chance that the answer may be wrong. A parameter that is not always produced is the reliability of a given set of identifications, which is of particular importance when individual datasets are to be compared, or when they are used to design additional (biological) studies.

An efficient way to estimate the false positive rate in a dataset is the use of reverse database.[28−30] Such a database is composed of proteins in normal (forward) direction, complemented with the same proteins with their sequences reversed (C- to N-terminus). The number of proteins identified from reverse proteins relative to the total number of identifications can then be used to calculate the false positive rate.

Being able to calculate false positive rates is one thing, the ability to minimize them is yet another. The only way to do so is by using other parameters than database scores only. Peptide retention time during chromatographic separation is one such a parameter, the isoelectric point of peptides, introduced recently by others,[19,22,31] is another.

In this study, we have investigated whether the overall false positive rate in a large-scale experiment can be decreased by peptide IEF followed by elimination of outliers deviating from the expected pI.

First, we have investigated the behavior of peptides during in-gel IEF using IPG strips since IPG protocols were initially developed for proteins.[7] Our data show that IEF of peptides is a relatively fast process, with focusing times that are around 10 times shorter than for proteins (3.5 kVh vs typically 32 kVh). This is most likely due to the much smaller size of peptides that migrate more easily through the gel than proteins. However, peptides with a basic p*I* require longer running times and/or higher focusing voltages than acidic peptides so that the running conditions of an in-gel peptide IEF may directly depend on the chosen pH range of the IPG strip. We have also noticed that, in contrast to protein IEF, peptide IEF is more tolerant to salt. Although under such conditions high voltages are not reached, efficient focusing is still achieved (data not shown).

A second important observation was that peptides can be accurately focused using IPG gel strips. This has allowed us to set boundaries excluding those peptides with outlying p*I*. Most of the peptides retained after filtering are confined to a single or at most two gel sections (Figure 4), which is a remarkable observation given the broad pH range allowed per section and the considerable overlap between them (Supporting Information Table 2). For instance, a peptide with pI 6.0 could theoretically be found anywhere between fractions 8 and 15 (Supporting Information Table 2). The observation that this is clearly not the case for the large majority of the peptides probably illustrates that the algorithm used to calculate peptide p*I* is inaccurate, and that the observed spread in p*I* in each fraction (1 and 0.4 pH unit before and after filtering, respectively, Supporting Information Figure 1) is still a considerable underestimation of the resolving power of IEF. The inaccuracy can be due to the fact that the algorithm was initially developed for proteins, and not peptides. Furthermore, the specific conditions during IEF (8 M urea) could effect electrostatic interactions influencing p*I*. An improved algorithm to more precisely predict p*I* of peptides would further enhance the potential of p*I* filtering.

Elimination of p*I* outliers after peptide IEF and FT−MS analysis resulted in the decrease of false positive identifications from 2.7% to 2.0%. The modesty of this improvement may in fact tell more about the FT−MS than about the utility of p*I* filtering, since the FP rate was already very low before filtering. This was different when the same sample was analyzed using an LTQ iontrap, where FP rates were considerably higher (9.9%), most likely due to the lower mass accuracy of this instrument. Here, p*I* filtering reduced the FP rate substantially to 3.0%. This comparison of the two datasets shows two things. One is that determination of accurate precursor mass helps significantly to obtain high-confident peptide identifications with a minimal number of false positives. Second, with a filtering approach such as IEF available, it is not a necessity to

use a high-end instrument such as an FT-MS in order to obtain identifications with high confidence (although this would go at the cost of the number of identifications).

An important question in this study has also been the comparison of IEF to a more widely used tool as a first step in the analysis of complex mixtures of proteins, SDS-PAGE gel electrophoresis. In a pair wise comparison of these methods using the same amount of starting material, both FT−MS and LTQ−MS resulted in significantly more peptide identifications using IEF (Table 1): 22% more peptides were identified with FT-MS (27% unique peptides), 10% more when using LTQ−MS (16% unique peptides). This increased sensitivity of IEF relative to SDS-PAGE may be related to its resolving power: of the 9265 unique peptides that remained after FT mass spectrometry and p*I* filtering, 7679 were found only once, and 1290 twice (in adjacent fractions), together representing 96% of all peptides. This in itself reflects a unique feature of IEF setting it apart from other separation techniques (chromatography, electrophoresis) where peptides or proteins can be separated but not concentrated in confined fractions. In fact this may be one of the features contributing to the sensitivity of IEF compared to SDS-PAGE. While in IEF peptides were found in 6 fractions at maximum, in SDS-PAGE this was in up to 17 fractions also leading to a higher peptide redundancy. Similarly, in another study IEF was compared to SCX chromatography as a prefractionation step.[32] More peptides were identified after IEF, despite the fact that only peptides separated across one pH unit were considered.[32]

One of the aims of this study was to generate the best possible list of identifications in the *Drosophila* nuclear extract containing the largest possible number of proteins with the lowest possible false positive rate. For the compilation of such a data set we have chosen a FP rate of 0.5%, and have set the protein cutoff values accordingly. This is different from usual procedures where a fixed cutoff is chosen. Of course one can select a safe setting (e.g., 60), but very likely this goes at the cost of the number of (probably correctly identified) proteins. From the numbers that emerge from our study (lower cutoff in FT compared to LTQ data) it seems that the cutoff value that is allowed may mainly be influenced by the quality of MS spectra. Other factors suggested previously may be sample complexity and size of the database searched,[30] but both of these were constant in our study.

Another (drastic) way to reduce the FP rate may be to reject all proteins based on single-peptide identifications. This would have been a valid approach in our study as well, since by far most "reverse proteins" were identified by a single peptide. However, this would have gone at the expense of a large number of identifications, introducing a high false negative rate. For instance, in our IEF−FT dataset, 423 out of 1751 proteins were identified by 1 peptide. By choosing a FP rate of 0.5% the confidence in all protein identifications is reflected, including those based on a single peptide. However, in studies where no FP can be calculated, especially in data sets with relatively low mass accuracies, removal of 1-hit wonders can be an effective measure to raise the confidence.

Conclusion

It is a prerequisite of present-day proteomics efforts to produce protein identifications at a high-confidence level. In this study, we have verified two parameters that are a useful aid in delivering such data, mass accuracy, and p*I* filtering after peptide IEF. Using an FT mass spectrometer, proteins can be identified with a minimal number of false positive identifications, while IEF provides an excellent way to reduce FP rates when high mass accuracies cannot be obtained. It is thus possible to present high-confident data using lower end MS instrumentation such as ion traps.

Acknowledgment. We gratefully acknowledge Prof. Peter Verrijzer (ErasmusMC, Rotterdam, The Netherlands) for providing *Drosophila* nuclear extract, and Dr. Elisabeth Gasteiger (SIB, Geneva, Switzerland) for providing the algorithm for p*I* calculations. We thank Dr. Bas van Breukelen for bioinformatic support. This work was supported by The Netherlands Proteomics Centre.

Supporting Information Available: Supporting Information Table 1 lists all peptides identified in this study. Supporting Information Table 2 shows the cutoff values used for filtering based on p*I*. Supporting Information Tables 3 and 4 list all proteins and peptides, respectively, that were identified in this study with a false positive rate of 0.5%. Supporting Information Figures 1 and 2 show the result of p*I* filtering after IEF followed by FT−MS and LTQ−MS, respectively. Supporting Information Figure 3 shows the number of identified peptides after SDS-PAGE followed by FT−MS and LTQ−MS. Supporting Information Figure 4 shows the distribution of peptide length and p*I* after IEF and SDS-PAGE followed by FT−MS and LTQ−MS. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, 198−207.
(2) Patterson, S. D.; Aebersold, R. H. Proteomics: the first decade and beyond. *Nat. Genet.* **2003**, *33* Suppl, 311−323.
(3) de Hoog, C. L.; Mann, M. Proteomics. *Annu. Rev. Genomics Hum. Genet.* **2004**, *5*, 267−293.
(4) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. Orthogonality of separation in two-dimensional liquid chromatography. *Anal. Chem.* **2005**, *77*, 6426−6434.
(5) Stasyk, T.; Huber, L. A. Zooming in: fractionation strategies in proteomics. *Proteomics* **2004**, *4*, 3704−3716.
(6) Wang, H.; Hanash, S. Intact-protein based sample preparation strategies for proteome analysis in combination with mass spectrometry. *Mass Spectrom. Rev.* **2005**, *24*, 413−426.
(7) Gorg, A.; Weiss, W.; Dunn, M. J. Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **2004**, *4*, 3665−3685.
(8) Issaq, H. J.; Chan, K. C.; Janini, G. M.; Conrads, T. P.; Veenstra, T. D. Multidimensional separation of peptides for effective proteomic analysis. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **2005**, *817*, 35−47.
(9) 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.
(10) Peng, J.; Gygi, S. P. Proteomics: the move to mixtures. *J. Mass Spectrom.* **2001**, *36*, 1083−1091.
(11) Huber, L. A.; Pfaller, K.; Vietor, I. Organelle proteomics: implications for subcellular fractionation in proteomics. *Circ. Res.* **2003**, *92*, 962−968.
(12) Sadygov, R. G.; Cociorva, D.; Yates, J. R., 3rd. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **2004**, *1*, 195−202.
(13) Chamrad, D. C.; Korting, G.; Stuhler, K.; Meyer, H. E.; Klose, J.; Bluggel, M. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics* **2004**, *4*, 619−628.

(14) Kapp, E. A.; Schutz, F.; Connolly, L. M.; Chakel, J. A.; Meza, J. E.; Miller, C. A.; Fenyo, D.; Eng, J. K.; Adkins, J. N.; Omenn, G. S.; Simpson, R. J. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics* **2005**, *5*, 3475−3490.
(15) Chen, Y.; Kwon, S. W.; Kim, S. C.; Zhao, Y. Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. *J. Proteome Res.* **2005**, *4*, 998−1005.
(16) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383−5392.
(17) Sadygov, R. G.; Liu, H.; Yates, J. R. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal. Chem.* **2004**, *76*, 1664−1671.
(18) Strittmatter, E. F.; Kangas, L. J.; Petritis, K.; Mottaz, H. M.; Anderson, G. A.; Shen, Y.; Jacobs, J. M.; Camp, D. G., 2nd; Smith, R. D. 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.
(19) Tan, A.; Pashkova, A.; Zang, L.; Foret, F.; Karger, B. L. A miniaturized multichamber solution isoelectric focusing device for separation of protein digests. *Electrophoresis* **2002**, *23*, 3599−3607.
(20) Xie, H.; Bandhakavi, S.; Griffin, T. J. Evaluating preparative isoelectric focusing of complex peptide mixtures for tandem mass spectrometry-based proteomics: a case study in profiling chromatin-enriched subcellular fractions in *Saccharomyces cerevisiae*. *Anal. Chem.* **2005**, *77*, 3198−3207.
(21) Chen, J.; Balgley, B. M.; DeVoe, D. L.; Lee, C. S. Capillary isoelectric focusing-based multidimensional concentration/separation platform for proteome analysis. *Anal. Chem.* **2003**, *75*, 3145−3152.
(22) Cargile, B. J.; Talley, D. L.; Stephenson, J. L., Jr. Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pI predictability of peptides. *Electrophoresis* **2004**, *25*, 936−945.
(23) Cargile, B. J.; Bundy, J. L.; Freeman, T. W.; Stephenson, J. L., Jr. Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. *J. Proteome Res.* **2004**, *3*, 112−119.
(24) Cargile, B. J.; Sevinsky, J. R.; Essader, A. S.; Stephenson, J. L., Jr.; Bundy, J. L. Immobilized pH gradient isoelectric focusing as a first-dimension separation in shotgun proteomics. *J. Biomol. Tech.* **2005**, *16*, 181−189.
(25) Heberlein, U.; Tjian, R. Temporal pattern of alcohol dehydrogenase gene transcription reproduced by Drosophila stage-specific embryonic extracts. *Nature (London)* **1988**, *331*, 410−415.
(26) Rappsilber, J.; Ishihama, Y.; Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **2003**, *75*, 663−670.
(27) Krijgsveld, J.; Ketting, R. F.; Mahmoudi, T.; Johansen, J.; Artal-Sanz, M.; Verrijzer, C. P.; Plasterk, R. H.; Heck, A. J. Metabolic labeling of C. elegans and, D. melanogaster for quantitative proteomics. *Nat. Biotechnol.* **2003**, *21*, 927−931.
(28) Moore, R. E.; Young, M. K.; Lee, T. D. Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 378−386.
(29) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC−MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2003**, *2*, 43−50.
(30) Qian, W. J.; Liu, T.; Monroe, M. E.; Strittmatter, E. F.; Jacobs, J. M.; Kangas, L. J.; Petritis, K.; Camp, D. G., 2nd; Smith, R. D. Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. *J. Proteome Res.* **2005**, *4*, 53−62.
(31) Michel, P. E.; Reymond, F.; Arnaud, I. L.; Josserand, J.; Girault, H. H.; Rossier, J. S. Protein fractionation in a multicompartment device using Off-Gel isoelectric focusing. *Electrophoresis* **2003**, *24*, 3−11.
(32) Essader, A. S.; Cargile, B. J.; Bundy, J. L.; Stephenson, J. L., Jr. A comparison of immobilized pH gradient isoelectric focusing and strong-cation-exchange chromatography as a first dimension in shotgun proteomics. *Proteomics* **2005**, *5*, 24−34.

PR0601180