Optimization and Use of Peptide Mass Measurement Accuracy in Shotgun Proteomics

Wilhelm Haas‡, Brendan K. Faherty§, Scott A. Gerber‡, Joshua E. Elias‡, Sean A. Beausoleil‡, Corey E. Bakalarski‡, Xue Li‡, Judit Villén‡, and Steven P. Gygi§

Mass spectrometers that provide high mass accuracy such as FT-ICR instruments are increasingly used in proteomic studies. Although the importance of accurately determined molecular masses for the identification of biomolecules is generally accepted, its role in the analysis of shotgun proteomic data has not been thoroughly studied. To gain insight into this role, we used a hybrid linear quadrupole ion trap/FT-ICR (LTQ FT) mass spectrometer for LC-MS/MS analysis of a highly complex peptide mixture derived from a fraction of the yeast proteome. We applied three data-dependent MS/MS acquisition methods. The FT-ICR part of the hybrid mass spectrometer was either not exploited, used only for survey MS scans, or also used for acquiring selected ion monitoring scans to optimize mass accuracy. MS/MS data were assigned with the SEQUEST algorithm, and peptide identifications were validated by estimating the number of incorrect assignments using the composite target/decoy database search strategy. We developed a simple mass calibration strategy exploiting polydimethylcyclosiloxane background ions as calibrant ions. This strategy allowed us to substantially improve mass accuracy without reducing the number of MS/MS spectra acquired in an LC-MS/MS run. The benefits of high mass accuracy were greatest for assigning MS/MS spectra with low signal-to-noise ratios and for assigning phosphopeptides. Confident peptide identification rates from these data sets could be doubled by the use of mass accuracy information. MS/MS spectra with low signal-to-noise ratios and for assigning phosphopeptides. Confident peptide identification rates from these data sets could be doubled by the use of mass accuracy information. The protein composition of the analyzed sample is ultimately inferred from identified peptides.

Recent MS technologies facilitate the acquisition of many thousands of MS/MS spectra over the course of one LC-MS/MS analysis. In general, only a portion of these spectra are correctly assigned to peptide ions using current database search algorithms. Low quality MS/MS data, non-peptide ions selected for fragmentation, and peptide ions where the corresponding sequence is not predicted from the searched database may contribute to this phenomenon. Commonly scores reflecting the quality of the match of acquired MS/MS data to predicted spectra are used as filter criteria to remove incorrect assignments. Manual validation of MS/MS spectra can help distinguish between true and false assignments but becomes infeasible with the ever increasing size of data sets. Only recently have statistical tools been developed that support the validation of peptide and protein assignments. Furthermore extending the information from mass spectroscopic data by using multiple MS stages (MSn) (11, 12) or by applying alternative fragmentation techniques extending the information for the commonly used collision-induced dissociation (CID) (13) was proposed for increasing the specificity of the peptide identification process. Also new scoring algorithms incorporating additional information from common CID-MS/MS spectra such as fragment ion intensities have been presented as useful tools for improved differentiation between correct and false peptide assignments.

The confidence in the identification of a peptide can also be enhanced by accurately measuring the mass of the peptide.


MS has become the method of choice for the characterization of complex protein mixtures from cells or tissues (1). The highest throughput and most comprehensive efforts to catalog protein mixtures have so far been achieved using a strategy known as shotgun proteomics (2). Proteins are proteolytically digested, and the resultant peptide mixture is usually separated by reversed-phase LC coupled on line to a mass spectrometer. Peptides are ionized and subjected to sequencing by MS/MS. Peptide ion sequences are derived based on fragment ions predominately formed through backbone cleavage at amide bonds. Several algorithms such as SEQUEST (3) and Mascot (4) allow automated assignment of MS/MS spectra by matching acquired data with spectra predicted on the basis of protein sequence databases (5). The protein composition of the analyzed sample is ultimately inferred from identified peptides (6).
ion. The importance of high mass accuracy in characterizing peptides has been described in numerous studies (16–21). Mass spectrometers that provide high mass accuracy include TOF (22) and FT-ICR (23) instruments. Their increasing usage in shotgun proteomic studies demands a thorough study of the role of high mass accuracy in such studies where peptide identifications were until recently based primarily on the rich information given by MS/MS spectra.

To accomplish this task, we used a hybrid linear quadrupole ion trap/FT-ICR (LTQ FT) mass spectrometer (Thermo Electron) (24) for the characterization of a complex peptide mixture. FT-ICR instruments provide the highest mass accuracy of the present MS technologies (for reviews, see Refs. 23 and 25), but challenges in their use when coupled on line to separation techniques restricted their application in large scale proteomic experiments. Foremost among these challenges were relatively large time scales for acquiring MS/MS data as well as practical limitations in achieving high mass accuracy based on the highly variable production of ions across a chromatographic separation. These difficulties have been addressed with the development of the hybrid LTQ FT mass spectrometer because the linear ion trap (26) allows high speed performance of MS/MS experiments and the adjustment of the number of ions analyzed in the ICR cell through automatic gain control (AGC).  

In the present study, we used a fraction of yeast whole-cell lysate tryptic digest as a model complex peptide mixture and analyzed the sample using different data-dependent MS/MS acquisition strategies. To ensure a high acquisition rate, the FT-ICR mass spectrometer was used only for accurate determination of peptide masses, whereas MS/MS experiments were entirely performed in the linear ion trap. The different acquisition strategies produced data sets providing peptide mass accuracies in either the low ppm range or a much wider range typical of traditional ion trap mass spectrometers often used in proteomic experiments. We applied the composite target/decoy database approach to validate the peptide assignments achieved from these data sets (8, 10). Ensuring a similar false-positive rate of the identified peptides allowed a fair comparison of the MS/MS data acquisition methods used and the role of mass accuracy in the peptide identification process. We also developed a simple mass calibration strategy for FT-ICR MS data from shotgun proteomic experiments that helped to avoid compromises between mass accuracy and the number of MS/MS spectra acquired in an LC-MS/MS run. The calibration procedure included the exploitation of polydimethylcyclosiloxane ions commonly detected as background ions in microcapillary LC-MS experiments as calibrant ions (27). To extend our understanding of the role of mass accuracy in the peptide profiling of samples different from those analyzed in the present study we artificially varied the size of the protein sequence database used for MS/MS spectra assignment as well as the quality of the MS/MS data. We also studied the role of peptide mass accuracy in the interpretation of MS/MS spectra from a 100-fold dilution of the starting sample and from phosphopeptides.

**EXPERIMENTAL PROCEDURES**

*Saccharomyces cerevisiae Growth Conditions, Lysis, Lysate Fractionation, Protein Digestion, and Enrichment of Phosphopeptides Using Immobilized Metal Affinity Chromatography—S. cerevisiae* (strain DMY1737) was grown and lysed as described previously (28). The protein content of the cleared lysate was determined using a Bradford protein assay (Bio-Rad) with BSA as an external standard. Disulfide bonds were reduced by adding DTT to a final concentration of 3 mM and incubating the solution at 56 °C followed by derivatization of cysteine residues with iodoacetamide (−5 mM excess in relation to concentration of reducing agents, 2-mercaptoethanol and DTT) at room temperature in the dark for 20 min. SDS, glycerol, and bromphenol blue were added to the protein solution to a total concentration of 2, 10, and 0.1%, respectively, and 1 mg of protein was loaded on a hand-poured preparative 10% polyacrylamide gel (2.6% bisacrylamide, Bio-Rad). Following electrophoresis and staining with colloidal Coomassie (Pierce) proteins of a molecular mass of 60–110 kDa were subjected to in-gel digestion with trypsin (modified sequencing grade porcine trypsin, Promega, Madison, WI) as described previously (29). Briefly cutting the corresponding gel region into 1-mm³ cubes was followed by destaining with 50 mM NH₄HCO₃, 50% acetonitrile; dehydration of the gel pieces; 45-min incubation with a solution of 12.5 μg/ml trypsin in 50 mM NH₄HCO₃ on ice; overnight digestion at 37 °C; and extraction of the peptides with 50% acetonitrile, 5% formic acid. Peptides were subjected to C₁₈ solid phase extraction (Vydac, Hesperia, CA) and dissolved in 5% ACN, 5% FA. For preparing a phosphopeptide sample 200 μg of the described reduced and alkylated yeast lysate were separated by SDS gel electrophoresis, and the digest of proteins of a molecular mass of 80–120 kDa was subjected to immobilized metal affinity chromatography using PHOS-Select iron affinity gel (Sigma) following the manufacturer’s instructions.

**Nanoscale Microcapillary Liquid Chromatography Electrospray Ionization Tandem Mass Spectrometry—LC-MS/MS experiments were performed in triplicate on an LTQ FT mass spectrometer (Thermo Electron, San Jose, CA) equipped with a Finnigan Nanospray II electrospray ionization source (Thermo Electron), an Agilent 1100 Series binary HPLC pump (Agilent Technologies, Palo Alto, CA), and a Famos autosampler (LC Packings, San Francisco, CA). Peptide mixtures were separated on a fused silica microcapillary column with an internal diameter of 125 μm and an in-house prepared needle tip with an internal diameter of ∼5 μm. Columns were packed to a length of 18 cm with a C₁₈ reversed-phase resin (Magic C18AQ; particle size, 5 μm; pore size, 200 Å; Michrom Bioresources, Auburn, CA). 4 μl of sample solution (−1 μg/μl, or 0.01 μg/μl for dilution experiments) were loaded onto the column, and separation was achieved by using a mobile phase from 2.5% ACN, 0.15% FA (buffer A) and 97.5% ACN, 0.15% FA (buffer B) and applying a linear gradient from 3 to 37% buffer B for 90 min (60 min for the analysis of the phosphopeptide sample) at a flow rate of 300 nl/min provided across a flow splitter by the HPLC pumps. An electrospray voltage of 2.1 kV was applied via a gold electrode through a polyetheretherketone junction at the inlet of the microcapillary column.

---

1 The abbreviations used are: AGC, automatic gain control; ΔCn, delta cross-correlation; CAI, codon adaptation index; FA, formic acid; FP, false-positive (peptide/protein assignment/identification); SIM, selected ion monitoring; S/N, signal-to-noise ratio; TIC, total ion current; TP, true-positive (peptide/protein assignment/identification); XCorr, cross-correlation score; LTQ, linear quadrupole ion trap.
The LTQ FT mass spectrometer was operated in the data-dependent mode using three different acquisition strategies as follows. With the SIM3 method (21) (see Fig. 1A) a scan cycle was initiated with a full-scan survey MS experiment (m/z 350–1700) performed with the FT-ICR mass spectrometer. The three most abundant ions detected in this scan were subjected to an FT-ICR selected ion monitoring (SIM) scan followed by an MS/MS experiment in the linear quadrupole ion trap (LTQ) mass spectrometer. Accumulation of ions for both MS and MS/MS scans was performed in the linear ion trap, and the AGC target values were set to 1 × 10^6 ions for survey MS, 5 × 10^6 ions for SIM, and 1 × 10^6 ions for MS/MS experiments. The maximum ion accumulation time was 250 and 150 ms in the FT-ICR and MS/MS modes, respectively. The resolution for the survey MS and to 5 × 10^6 for the SIM scans. Isolation widths were ± 5 m/z for SIM and ± 1.25 m/z for MS/MS experiments, and ions were selected for MS/MS when their intensity exceeded a minimum threshold of 1200 counts. Singly charged ions were not subjected to MS/MS. The normalized collision energy was set to 35%, and one microscan was acquired per spectrum. Ions subjected to MS/MS were excluded from further sequencing for 30 s.

When using the FT10 method (see Fig. 2A) the acquisition of SIM scans was omitted. A full-scan survey MS experiment (m/z range as above; AGC target, 1 × 10^6 ions; resolution, 1 × 10^6; maximum ion accumulation time, 1000 ms) was acquired in the FT-ICR mass spectrometer followed by MS/MS experiments on the 10 most abundant ions detected in the full-MS scan. MS/MS spectra were collected in the LTQ mass spectrometer, and the settings were identical to those described for the SIM3 method.

With the LTQ10 method only the LTQ and not the FT-ICR mass spectrometer was used (Supplemental Fig. 3). A full-scan MS was again followed by 10 MS/MS experiments on the 10 most abundant ions detected in the full-MS MS. The survey MS AGC target was set to 2 × 10^6 ions, and its maximum ion accumulation time was 150 ms. MS/MS settings were as described above with the exception that MS/MS was also performed on singly charged peptides.

**Data Processing, Database Searching, and Mass Calibration Procedures**—Instrument control and primary data processing were done using the Xcalibur software package, Version 1.4 SR1 (Thermo Electron). Data were originally stored in the .RAW format. LTQ10 MS/MS data including no information from FT-ICR data were converted into the .dta format, the input data format for SEQUEST searches (see below), by the program ExtractMS Version 2.11 (Thermo Electron, fields.scripps.edu/sequest/extractms.html), which separates singly charged peptide MS/MS spectra from those of multiply charged peptide. As the unit mass resolution of MS data acquired with the LTQ mass spectrometer did not allow the determination of the charge state of multiple charged peptide ions, three .dta files were created for each corresponding MS/MS spectrum assuming potential charge states of +2, +3, and +4. Data from analyses including the use of the FT-ICR mass spectrometer were extracted into the OpenRaw format using the program xr2or written in Visual C++ (downloaded from clu.bmed.harvard.edu/MapQuant/) (30). Based on this data structure, in-house Perl scripts were used to extract the exact measured monoisotopic m/z as well as the charge state of peptide ions selected for MS/MS experiments, and together with MS/MS fragment ion m/z and intensity data, this information was included into files created in the .dta format for the acquired MS/MS spectra (31).

Using SEQUEST (Version 27, Revision 12) on a Linux cluster with 17 dual 1.5–2.4-GHz processor nodes or the SEQUEST Sorcerer platform (www.sagenresearch.com), MS/MS data in the .dta format were searched against a composite target/decoy protein sequence database in which the target component was comprised of sequences derived from the known S. cerevisiae ORFs (downloaded September 10, 2004 from the Saccharomyces Genome Database (SGD) at Stanford University, ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/orf_protein/) and protein sequences of known contaminant proteins, such as porcine trypsin and human keratins (6427 entries total). This component was followed in the database by a decoy component composed of the reversed sequences of all proteins in the target component. For simulating the search of the acquired MS/MS spectra against a larger database, the size of the yeast ORF protein sequence database was extended by a factor of nine (57,851 entries) using a Markov chain model (fourth order) based on the protein sequences in the original database. A target/decoy version of this database was created as described above. The signal-to-noise ratio of each MS/MS spectra was modified by increasing the intensity of the lowest 50% of signals from the original spectra by a factor of 30. Both database and data set manipulations were accomplished with in-house Perl scripts.

Database searches were performed by applying precursor ion m/z tolerances of 3 ppm (SIM3), 6 ppm (FT10), or ±2 Da (LTQ10) and fragment ion m/z tolerances of ±1 Da. Cysteine residues were searched as carbamidomethylated (mass increment of 57.02146 Da), and methionine residues were allowed to be oxidized (+15.99492 Da). When assigning MS/MS spectra from phosphopeptide analyses, serine, threonine, and tyrosine residues were allowed to be phosphorylated (+79.96633 Da). For the search of phosphopeptide MS/MS data only tryptic peptides were considered for matching with the acquired MS/MS spectra in the database search. For the assignment of MS/MS spectra from unmodified peptides no enzyme specificity constraints were applied, and peptide assignments were only accepted when both peptide termini were consistent with trypsin specificity. Three assignments were made for most LTQ10 MS/MS spectra as they were considered as doubly, triply, or quadruply charged peptides (see above). Two assignments for each spectrum were removed from the data set before further data analysis. For nonspecific searches, this was done by using the trypic state of the peptide as a primary filter criterion (in the presence of trypic assignments these were preferred over non-tryptic ones) and the cross-correlation score (XCorr) (see below) of the assignment as secondary criteria (primary for trypic searches).

False-positive (FP) peptide assignments were removed by filtering on the basis of two metrics calculated by the SEQUEST program that express the quality of the match between an experimental and a database-predicted MS/MS spectrum. The first was the XCorr, which reflects the quality of the match of the two spectra. The second was the delta cross-correlation (ΔCn), which is the normalized difference between the XCorr values for the two best peptide sequence matches to an acquired MS/MS spectrum (3, 5). XCorr and ΔCn thresholds were applied to obtain a peptide identification FP rate of ~1%. The FP rate was estimated using the target/decoy database approach: assuming true-positive (TP) peptide identifications were exclusively assigned to peptides from the target database component and FP identifications were equally distributed between target and decoy components, the number of FP peptide identifications was estimated by doubling the number of decoy database assignments and dividing the result by the number of total peptide identifications in the data set (10). The fraction of FP identifications was expressed as a percentage (FP rate). The estimation of protein identification false-positive rates was based on the same calculation (for examples see Supplemental Figs. 4 and 5). XCorr and ΔCn filtering were applied separately for assignments to doubly, triply, and quadruply charged peptide ions. Assignments to singly charged peptide ions as well as to peptide ions with fewer than eight amino acids were not considered in this study and removed from the data set before further analysis. When the.data
from triplicate analyses were studied, XCorr and ΔCn thresholds required to achieve a 1% peptide identification FP rate were determined on the basis of combined data sets.

Initial external mass calibration for the LTQ FT mass spectrometer was done as recommended by the manufacturer using singly protonated ions of caffeine, a peptide with the sequence MRFA, and Ultramark 1621 (a mixture of fluorinated phosphazenes). Recalibration of FT-ICR peptide ion m/z values was done by using the following calibration function,

\[ m/z = a/f + b/f^2 + c \text{TIC}/f^2 \]  
(Eq. 1)

where \( f \) is the detected cyclotron frequency; \( a, b, \) and \( c \) are calibration coefficients; and TIC is the total ion current measured for an MS spectrum (33–36). For recalibration, calibration coefficients recorded in the .RAW file format were used to recalculate \( f \) from \( m/z \) values using the calibration function described previously (33). TIC values stored in the .RAW format had been normalized accounting for varying ion accumulation times during an analysis. For use in Equation 1, this normalization was reversed by multiplying the TIC values by the ion accumulation time in seconds. Singly protonated polydimethylsiloxane ions, \((\text{Si(\text{CH}_3)_2O})_n\), which are commonly detected as background ions in nanoscale LC-MS/MS experiments (27) were exploited as calibrants. Ions containing five \((m/z\ 371.101237)\), six \((m/z\ 445.120029)\), seven \((m/z\ 519.138821)\), eight \((m/z\ 593.157613)\), and nine \((m/z\ 667.176405)\) dimethylsiloxane monomers were used when their intensity exceeded 1000 counts. The Levenberg-Marquardt algorithm implemented in the SigmaPlot software package, Version 9 (Systat Software, Point Richmond, CA) was used for optimizing calibration coefficients for the calibration function (Equation 1). Peptide ion signal-to-noise ratios (S/N) were determined using an in-house developed software package for automated peptide quantification on the basis of MS data (VISTA)\(^3\).

**RESULTS**

**Optimization of Mass Accuracy on the LTQ FT Mass Spectrometer**—The instrument platform used in this study was an LTQ FT mass spectrometer (Thermo Electron) (24). FT-ICR mass spectrometers provide the highest mass accuracy of the present MS technologies (23), but there are practical limitations to the performance of these instruments. In FT-ICR MS, ion m/z values are determined based on the frequency of the cyclotron motion that ions perform under the influence of a magnetic field in the ICR cell. However, this frequency is also affected by electric fields including those produced by the analyzed ions themselves. These phenomena are termed space-charge effects, and they evoke errors in mass measurement in relationship to the number of ions analyzed in the mass spectrometer. Although various methods have been developed to correct for these errors (25), space-charge effects are still a substantial problem especially when FT-ICR MS is coupled to separation techniques where highly variable numbers of ions may be introduced into the mass spectrometer over the course of an analysis. With the development of the LTQ FT mass spectrometer, this problem has been almost completely addressed as the linear trap allows for fine control over the number of ions analyzed in the ICR cell. This feature is termed AGC and has recently been exploited by Olsen et al. (21) to optimize the mass accuracy for LTQ FT large scale proteomic applications. The described method involved the determination of peptide ion m/z values by FT-ICR SIM scans recorded over a narrow m/z range around the analyzed ions. The small m/z range allowed high peptide ion S/N values even when only a small number of ions were analyzed in the ICR cell (37). By using the AGC to ensure that a constantly low number of ions was subjected to SIM scans, space-charge effects were avoided, and an excellent mass accuracy was achieved. An absolute mass error of less than 2 ppm was reported for 163 peptides identified from the protein carbamoyl-phosphate synthase identified in a high protein molecular weight fraction of a mouse liver proteome tryptic digest (21).

We applied the described method, here denoted as SIM3, to analyze a complex peptide mixture produced by using trypsin to digest a fraction of yeast whole-cell lysate. After recalibration of peptide ion masses (see below) MS/MS data

---

were searched against a yeast protein sequence database using the SEQUEST algorithm. The mass accuracy distribution for the identified peptide ions is shown in Fig. 1B. We were able to confirm the excellent performance of the SIM3 method. The average measured peptide mass error was $-0.41 \text{ ppm}$, and the mass accuracy distribution showed an S.D. of 0.44 ppm.

Although providing high mass accuracy, we found the acquisition of SIM scans to be time-consuming (Fig. 1A). Each MS/MS spectrum acquired in an LC-MS/MS peptide profiling experiment might potentially lead to the identification of a peptide. We therefore intended to improve the LTQ FT MS/MS acquisition rate by analyzing the sample with the FT10 method (Fig. 2). By omitting the recording of SIM scans the number of acquired MS/MS spectra was substantially increased. Within an average cycle time of $\sim 3.5 \text{ s}$, 10 MS/MS spectra were recorded in contrast to only three in a slightly shorter cycle time of 2.6 s with the SIM3 method. However,
with the FT10 method, peptide ion m/z values had to be determined from the survey MS scan where, in comparison with the SIM scans, an immensely higher number of ions (1,000,000) were analyzed. These scans provided an average peptide ion mass accuracy of 4.99 ± 2.42 ppm (Fig. 2C, panel 3), which was lower than that achieved with the SIM3 method. However, we considered the higher MS/MS acquisition rate of the FT10 strategy as a potentially critical feature for large scale proteomic studies. To avoid a compromise between acquisition rate and mass accuracy we sought to improve the accuracy provided by this method through mass recalibration of the acquired data.

A prerequisite for mass calibration is the presence of ions with known masses to be used as calibrants. Singly protonated polydimethylcyclosiloxanes, a group of air contaminants, are well known background ions commonly detected in nanoscale LC-MS experiments (27). We exploited five highly abundant species of these ions with m/z values of 371, 445, 519, 593, and 667 (see “Experimental Procedures” for exact values) as calibrants for postacquisition mass calibration for LTQ FT MS data. Polydimethylcyclosiloxane ions were detected above background noise level in only a portion of the ~1500 FT-ICR survey MS spectra acquired in an FT10 run (Supplemental Fig. 1B); this was ascribed to ionization suppression effects. To amend mass measurement errors in all MS spectra an external calibration strategy had to be applied where MS data from spectra showing no calibrant ion signals were corrected based on signals observed in other spectra. We modified a widely used FT-ICR MS calibration equation developed by Ledford et al. (33) to address differences in the number of ions analyzed in individual MS survey spectra during an FT10 analysis (see “Experimental Procedures” and Supplemental Fig. 1C) (34–36). By using this calibration strategy, we substantially improved the mass accuracy of FT10 data. When compared with uncalibrated data, the average mass deviation was reduced from 4.99 to −0.25 ppm, and the S.D. of the mass accuracy distribution from 2.42 to 1.46 ppm (Fig. 2C, panels 3 and 4). We observed that the peptide mass errors could be approximately fitted with a Gaussian distribution (Fig. 2). Thus, a minimum peptide mass search tolerance of three standard deviations of the mass accuracy distribution was required in an MS/MS spectra database search to allow correct assignment of more than 99% of the acquired spectra. Recalibration of FT10 survey MS data lowered the minimum peptide tolerance requirement for a database search of the MS/MS data from ~13 ppm to about 5 ppm. The relative high tolerance for the uncalibrated data resulted in part from the fact that the SEQUEST program does not allow adjustments for a shift in mass accuracy distributions. However, the overall mass accuracy of −0.25 ± 1.46 ppm of the recalibrated FT10 data was still worse than the −0.41 ± 0.44 ppm achieved with the SIM3 method. It has to be noted here that the SIM3 mass accuracy distribution displayed in Fig. 1B was also achieved by mass recalibration. About 5% of the SIM scans included polydimethylcyclosiloxane background ions that were exploited applying the mass calibration procedure described above. The mass accuracy distribution for the uncalibrated data showed a similar S.D., but an average mass accuracy shifted to 1.6 ppm (Supplemental Fig. 2).

The comparison of the SIM3 and the FT10 methods showed that the latter provided a slightly reduced mass accuracy but at the same time allowed more MS/MS spectra to be acquired in an LC-MS/MS run. Therefore, we next sought to evaluate the role of these two features, mass accuracy and MS/MS acquisition rate, in mass spectrometry-based proteomic studies by performing a detailed comparison of MS/MS data acquired with both methods.

**Benefits and Costs of High Mass Measurement Accuracy in Shotgun Proteomic Experiments**—In shotgun proteomic experiments, method-dependent differences in peptide mass accuracy achievable with the LTQ FT spectrometer present a challenge in determining the optimum peptide mass search tolerance for database searching of the acquired MS/MS data. Although narrowing the tolerance was reported to enhance the peptide identification process by decreasing the number of peptides that can be matched to the acquired MS/MS data (21), undershooting the mass accuracy provided by the mass spectrometer would prevent correct assignment of spectra. The smallest applicable tolerance may be determined by searching the data twice. Starting with a high mass tolerance, the settings for the second search would be adjusted in response to the results from the first. However, this strategy substantially increases data analysis time. We also observed that setting the tolerance based on results from data sets acquired with the same method is complicated by slight run-to-run differences in mass accuracy (data not shown). We addressed this problem by exploiting the above described external calibration procedure with polydimethylcyclosiloxane as calibrant ions. It was found that the mass accuracy distribution of background ions closely resembled that of peptide ions (Fig. 2C). Thus, after extracting background ions from the acquired MS survey or SIM spectra and mass recalibration of the MS data, the S.D. of the error observed for the recalibrated background ion m/z values can be used for defining the peptide mass search tolerance in subsequent database searching of MS/MS data (Fig. 2B). As we observed slightly narrower mass accuracy distributions for background ions than for peptide ions, we used five standard deviations of background ion distributions as peptide mass search tolerance. The highest values for this S.D. in three runs acquired in this study were 1.20 ppm for FT10 and 0.63 ppm for SIM3 analyses. Therefore, we used 6 and 3 ppm as peptide mass search tolerances for a SEQUEST database search of MS/MS spectra acquired with the FT10 and the SIM3 methods, respectively. Searches were done against a composite target/decoy yeast protein sequence database, which allowed the estimation of incorrect assignments in the final data sets as described under “Experimental Procedures” (8, 10).
High Mass Accuracy in Shotgun Proteomics

Number of MS/MS spectra, confident peptide identifications, and inferred protein identification from analyzing a fraction of yeast proteome digest using three different data-dependent MS/MS acquisition strategies in 90-min LC-MS/MS analyses (mean and S.D. from three analyses)

Peptide and protein identification false-positive rates are given in parentheses. IDs, identifications.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acquired MS/MS spectra</th>
<th>Peptide IDs</th>
<th>Protein IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTQ10</td>
<td>17,627 ± 573</td>
<td>3,663 ± 29</td>
<td>394 ± 7</td>
</tr>
<tr>
<td>FT10</td>
<td>14,701 ± 107</td>
<td>4,034 ± 36</td>
<td>397 ± 5</td>
</tr>
<tr>
<td>SIM3</td>
<td>5,920 ± 37</td>
<td>2,705 ± 28</td>
<td>236 ± 6</td>
</tr>
</tbody>
</table>

searched without using enzyme specificity constraints but filtered MS/MS spectra assignments for those to peptides with both termini consistent with trypsin specificity. When compared with matching MS/MS spectra only to fully tryptic peptides, this search and filter strategy was recently shown to increase the number of peptide identifications with a defined false-positive rate from SEQUEST database searches of MS/MS spectra from tryptic digests (38). We confirmed these findings in the present study (data not shown). We applied XCorr and ΔCn cutoffs (see Supplemental Table 1) to remove incorrect assignments from the filtered tryptic data set such that the FP identification rate in the final data set was ~1%. Peptide assignments with this FP rate are defined here as confident peptide identifications. The closely identical FP rates in the final peptide identification data sets from both methods, FT10 and SIM3, allowed a fair comparison of the two MS/MS acquisition strategies.

To study the effect of high mass accuracy on the peptide identification process from MS/MS spectra beyond the small differences observed for the accuracy achieved with the FT10 and the LTQ SIM3 TOP methods, we acquired a third MS/MS data set. The sample was analyzed in triplicate using the LTQ10 method. This produced a control data set resembling the quality of MS and MS/MS typical for ion trap mass spectrometers often used in large scale proteomic studies. The workflow of this method is depicted in Supplemental Fig. 3. By omitting the acquisition of FT-ICR mass spectra the cycle time was slightly reduced to 3 s when compared with that observed for the FT10 method. LTQ10 MS/MS data were assigned with the SEQUEST algorithm in a fashion similar to that described for the FT10 and the SIM3 data sets. The peptide mass tolerance was set to ±2 Da, which is a typical value for the database searches of ion trap MS/MS data. XCorr and ΔCn thresholds ensured an FP rate of ~1% for tryptic peptide assignments obtained from a search with no enzyme specificity constraints (Supplemental Table 1).

Table I and Fig. 3 display average numbers of confident peptide assignments and inferred protein identifications from triplicate analyses of the studied yeast proteome fraction by each of the described LC-MS/MS methods. Strikingly, the FT10 data set including high mass accuracy information produced only ~10% more confident peptide identifications than the LTQ10 data set acquired using only the LTQ mass spectrometer. The number of protein identifications was nearly identical for both methods. The SIM3 data set provided the highest mass accuracy but gave the smallest number (30% fewer peptide and 40% fewer protein identifications) compared with the other two methods. This is best explained by the substantially reduced number of MS/MS spectra acquired with the SIM3 method compared with that produced with the other two methods (Table I and Fig. 3A). SIM scans acquired with the SIM3 experiment reduced the number of MS/MS experiments relative to the LTQ10 and the FT10 methods by 70 and 60%, respectively (Table I). Fig. 3B shows that peptide ions missed with the SIM3 method were mainly from proteins predicted to be of low abundance as they were encoded by genes with codon adaptation index (CAI) values of 0.2 and lower. In addition, a generally lower number of unique peptides identified for a given protein (Fig. 3C) accompanied this effect. We sought for a more detailed analysis of both phenomena by studying peptide ion S/N as well as CAI of genes encoding for the corresponding proteins for peptides missed with the SIM3 method. We extracted both values for peptides identified with the FT10 method and plotted their relationship for ions identified with both methods, FT10 and SIM3, and for ions correctly assigned only based on FT10 data (Fig. 4). It was observed that most of the peptide ions identified with both methods had a S/N of 10 or higher, whereas a high portion of ions identified only with the FT10 method showed a S/N of lower than 10 (Fig. 4, B and C). These low S/N values correlated well with a low CAI of corresponding genes for the derived proteins. The FT10 method also allowed increased sequence coverage for identified proteins by confident identification of peptides producing only small ion signals (Fig. 4D).

The described costs of sacrificing a high MS/MS acquisition rate for obtaining a slightly enhanced mass accuracy were less surprising than the small general benefits high mass accuracy provided in the assignments of the acquired MS/MS spectra. FT10 data produced only 10% more confident peptide identifications than those acquired with the LTQ10 method. Furthermore these additional peptide assignments primarily increased the sequence coverage for the inferred protein identifications but not their number (Fig. 3). The peptide mass search tolerances applied in the assignment of MS/MS spectra were ±2 Da for LTQ10 and only 6 ppm for FT10 data. This difference reduced the mass range for peptides matched in a SEQUEST database search of an FT10 MS/MS spectrum for a 1000-Da peptide by a factor of ~300. The small benefit provided by the applied narrow peptide
mass search tolerance emphasizes the primary role of the information given by MS/MS fragment ions in the peptide identification process. It must be noted that the low resolution of LTQ survey MS data did not allow the determination of the charge states of peptide ions, and each MS/MS spectra of a potentially multiply charged peptide ion was considered here to be of either a doubly, triply, or quadruply charged ion.

Therefore, the number of searched MS/MS spectra was almost 3 times the number of acquired spectra. This not only greatly increased the database search time but also increased the number of misassigned MS/MS spectra as only one charge state could possibly be correct. However, the number of confidently assigned spectra showed that the tryptic state of the assigned peptides as well as XCorr scores (see “Experimental Procedures”) allowed effective discrimination of incorrect and correct assignments.

We next sought to estimate the role of high mass accuracy for the analysis of samples different from those analyzed in this study. Although yeast is a well established model organism for proteomic studies, the size of the yeast ORF protein sequence database is relatively small (6427 entries) when compared with databases for other species (57,478 entries in the human International Protein Index sequence database, September 5, 2005, www.ebi.ac.uk/IPI/IPIhuman.html). To simulate assignment of MS/MS spectra against a larger database, we artificially extended the number of entries in the original yeast protein sequence database by a factor of nine using a Markov chain model. LTQ10 and FT10 data were searched against the new database as described above. We also aimed to study the influence of the quality of MS/MS spectra on the role peptide mass accuracy has on their assignment. Therefore, we reduced the S/N of the acquired MS/MS data by multiplying the intensity of low abundance signals in the spectra by a factor of 30 before searching the manipulated data set against the original yeast protein sequence database. Because limited sample amount is an assumed reason for low S/N MS/MS spectra, we also analyzed approximately 0.04 μg of the above described fraction of yeast proteome tryptic digest in triplicate by LC-MS/MS using the FT10 and the LTQ10 methods. The analyzed sample amount was 1/100 of the original sample amount.

Results from these searches are summarized in Fig. 5 and Supplemental Table 2. Assigning MS/MS spectra using a larger database size slightly reduced the number of confident peptide identifications, a phenomenon that has been described previously (39). However, in comparison with the search against the original target/decoy database, the effect of high mass accuracy information did not substantially change. FT10 data produced only 12% more confident peptide identifications than LTQ10 data. Reducing MS/MS spectrum quality caused a large decrease in the number of identified peptides from both data sets, but the effect was more severe for the assignment of LTQ10 spectra. High mass accuracy information doubled the number of confidently assigned low quality MS/MS spectra. Analyzing a 100-fold decreased sample amount reduced the number of confidently identified peptides by a factor of only approximately two. Comparing LTQ10 and FT10 methods showed that accurately measured peptide masses generated 20% more confidently identified peptides. Considering the substantial decrease in analyzed sample amount, it was intriguing to observe a rela-
tive small differences in identified peptides (50%) compared with the original sample (Fig. 5D). We therefore analyzed another sample dilution of 25-fold by both methods with similar results (30% average decrease in identifications compared with original sample and only 10% difference in identified peptides by both methods; data not shown). These data suggest the AGC function of the LTQ FT mass spectrometer effectively compensated for the 100-fold decrease in the analyzed sample amount.

Examples of relatively noisy MS/MS spectra are those of phosphopeptides. These spectra are often dominated by ions resulting from the neutral loss of phosphoric acid, whereas sequence-specific fragment ions formed through cleavage of the peptide backbone amide bonds can be of low intensity showing a small S/N (11, 40). We sought to examine whether our results for the assignment of artificially altered MS/MS data also applied to phosphopeptide MS/MS spectra. We used IMAC to enrich for phosphorylated peptides (41, 42) from the digest of a fraction of the yeast proteome and analyzed the sample using 60-min LC-MS/MS methods with the FT10 and the LTQ10 data acquisition strategies. In the database search of the acquired MS/MS data, we allowed serine, threonine, and tyrosine residues to be phosphorylated. We found that applying no enzyme specificity for the peptide mass tolerance search of LTQ10 data followed by filtering for tryptic peptides lowered the number of confident peptide identifications when compared with a search only considering tryptic peptides (data not shown). The number of peptides potentially matching an MS/MS spectrum is greatly increased through allowing a variable modification on three different residues. This complicates the automated assignment of MS/MS spectra, especially when they may be of reduced quality (39). The nonspecific database search slightly increased the number of peptide assignments using a 6-ppm peptide mass tolerance for the acquired FT10 phosphopeptide MS/MS data. However, to use similar database search strategies we reduced the database size for the search of both data sets by considering only tryptic peptides.
cutoffs were used to remove false-positive assignments (Supplemental Table 1). Assignments to the same peptide modified at different positions usually show very similar XCorr values. ΔCn cutoffs would therefore remove a significant fraction of correct phosphopeptide identifications. Fig. 6 shows the results from the database searches at a 1% FP rate. Omitting the use of the FT-ICR part of the LTQ FT mass spectrometer produced 2422 MS/MS spectra with the LTQ10 method; this was substantially more than the 1642 spectra acquired with the FT10 method. The smaller number of acquired MS/MS spectra for these samples was ascribed to the fact that the phosphopeptide sample was less complex, available in a smaller amount than the above described yeast lysate digest, and only analyzed using a 1-h gradient. The high peptide mass accuracy provided by the FT10 method allowed the confident assignment of 396 phosphopeptide MS/MS spectra (1% FP rate). This was more than twice the number (190 spectra) assigned from the LTQ10 run. These results highlight the importance of mass accuracy for large scale phosphorylation studies.

**DISCUSSION**

MS/MS data are currently the main source of information for identifying large numbers of peptides generated by digestion of protein mixtures in so-called “bottom-up” mass spectrometry-based proteomic experiments. Recently mass spectrometers with the capability to provide exceptionally high mass accuracy and resolution of precursor ions along with rapid MS/MS collection capabilities have become available for widespread use (22, 24, 43). However, the measured mass accuracies are currently not sufficient to overcome the necessity of acquiring MS/MS information for the identification of peptides (18). This becomes even more apparent when one considers the as yet incompletely understood level of com-

---

**Fig. 5.** The benefits of high mass accuracy on the peptide identification process from different sample amounts, data set compositions, and database sizes. *A,* confident peptide identifications from FT10 and LTQ10 data at a 1% FP rate. MS/MS spectra were searched with the SEQUEST algorithm against a target/decoy composite database including all known yeast protein sequences. *B,* the same data sets were searched against a database generated by increasing the size of the yeast protein sequence database by a factor of nine using a Markov chain model (see “Experimental Procedures”). *C,* the S/N values from MS/MS spectra of the same data sets were artificially decreased, and the spectra were searched against the original yeast protein sequence database. *D,* the sample amount analyzed by both methods was decreased 100-fold, and the acquired MS/MS spectra were searched against the original yeast protein sequence database. Values are mean ± S.D. from triplicate analyses. *IDs,* identifications.

**Fig. 6.** The benefits of high mass accuracy for assigning phosphopeptide MS/MS spectra. Phosphopeptides were enriched from a fraction of yeast whole-cell lysate by IMAC (see “Experimental Procedures”). Identical sample amounts were analyzed using the FT10 and the LTQ10 strategies in 60-min LC-MS/MS runs. The numbers of acquired MS/MS spectra and confident phosphopeptide identifications from both data sets are shown. *IDs,* identifications.
plexity added to the proteome through post-translational modifications (44). However, accurately determined peptide masses are being increasingly used to support the automated assignment of MS/MS spectra to enhance the confidence in peptide identification (21, 45).

Our study showed complementary roles for peptide precursor ion mass accuracy and the quality of information provided by MS/MS data in the identification of peptide ions. We found that the benefit of obtaining accurate masses for precursor ions would appear to be greatly overestimated for the analysis of complex mixtures of unmodified peptides (Fig. 5A) even with highly diluted samples (Fig. 5D). Triplicate analyses of the yeast sample acquiring precursor ion masses with either the FT-ICR or an LTQ mass spectrometer gave surprisingly similar results (Fig. 5A). Although FT-ICR detection allowed a database search of the acquired MS/MS data with a peptide mass tolerance of 6 ppm while ± 2 Da was applied for searching LTQ data, the high mass accuracy only provided the confident identification of 10% more peptides and no observable difference in the number of protein identifications (Fig. 3A).

We did find, however, that the role of mass accuracy in the peptide identification process was largely increased for the assignment of MS/MS spectra with low quality. This was shown for the assignment of MS/MS data with artificially lowered S/N (Fig. 5C) and confirmed for the automated interpretation of phosphopeptide MS/MS spectra (Fig. 6). In the latter, b- and y-type ions often show low S/N as spectra from peptides phosphorylated at serine or threonine residues tend to be dominated by intense peaks from the neutral losses of phosphoric acid and water. For the interpretation of these MS/MS data, a 6-ppm peptide mass search space doubled the number of confidently identified peptides.

High mass accuracy data collected in the MS but not in the MS/MS mode support peptide identification through MS/MS database searches by reducing the number of peptides from the searched database that are potentially matching the acquired MS/MS data. It has been described previously that low quality spectra and large database sizes complicate the correct assignment of MS/MS data (39); this was confirmed in this study by showing the different roles of accurately determined peptide masses in the assignment of MS/MS data of different quality. Here we did not study the role of high mass accuracy in MS/MS spectra. Although accurately measured fragment ion masses are expected to considerably increase the certainties associated with each individual peptide identification, using the FT-ICR mass spectrometer for analyzing fragment ions in large scale proteomic experiments with the LTQ FT mass spectrometer substantially decreases the number of MS/MS spectra acquired in an LC-MS/MS run (13, 46).

Disadvantages of improving mass accuracy in an LC-MS/MS run at the cost of MS/MS acquisition rate were shown by a comparison of the SIM3 and the FT10 methods (Fig. 4). Although SIM scans provided a slightly better mass accuracy than the FT-ICR survey MS spectra acquired with the FT10 method, omitting SIM scans nearly tripled the number of acquired MS/MS scans resulting in a 1.5-fold increase for confident peptide and protein identifications (Fig. 3A).

We developed a simple mass recalibration procedure involving the use of commonly detected polydimethylclosiloxane background ions as calibrants. Using this procedure, we substantially improved the mass accuracy in FT-ICR survey MS scans; this avoided a compromise between mass accuracy and MS/MS acquisition rate in LTQ FT large scale proteomic experiments.

Acknowledgment—we thank D. Moazed, Department of Cell Biology, Harvard Medical School, for providing yeast lysate.

Addendum—After submission of this work the use of a polydimethylclosiloxane ion as calibrant ion for mass recalibration in proteomic studies was published in a study by Olsen et al. (32).

* This work was supported in part by National Institutes of Health Grants GM67945 and HG3456 (to S. P. G.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

¶ To whom correspondence should be addressed. Tel.: 617-432-3155; Fax: 617-432-1144; E-mail: steven_gygi@hms.harvard.edu.

REFERENCES
teomics by two consecutive stages of mass spectrometric fragmenta-
identification using complementary fragmentation techniques in Fourier
Intensity-based protein identification by machine learning from a library of
ments for peptide characterization by monoisotopic molecular mass
measurements. Anal. Chem. 68, 4060–4063
mass measurement (±10 ppm) in protein identification strategies em-
ploying MS or MS/MS and database searching. Anal. Chem. 71, 2871–2882
identification. Anal. Chem. 72, 3349–3354
D. L., Battista, J. R., Daly, M. J., Fredrickson, J. J., Kosan-
darthes, H., Masselon, C., Marklill, L. M., Moore, R. J., Romine, M. F.,
Shen, Y., Stritmatter, E., Tolic, N., Udseth, H. R., Venkateswaran, A.,
Deinococcus radiodurans proteome by using accurate mass tags. Proc.
A. G. (2004) Theoretical and experimental prospects for protein identifi-
cation based solely on accurate mass measurement. J. Proteome Res. 3, 61–67
C-terminal to arginine and lysine residues. Mol. Cell. Proteomics 3, 608–614
36, 849–865
Rev. 17, 1–35
J. C., Uebberheide, B., Garcia, B., Busby, S., Muratore, T., Shabanowitcz,
spectrometer: performance characterization and use in the comparative
analysis of histone H3 post-translational modifications. J. Proteome Res.
3, 621–626
mass measurements by Fourier transform mass spectrometer. Mass Spectrom.
Rev. 24, 286–309
loxanes in the ambient laboratory air identified as source of extreme
38, 523–525
RNAs associated with the RNA interference effector complex RITS.
Methods Enzymol. 392, 297–307
metric sequencing of proteins silver-stained polyacrylamide gels. Anal.
Chem. 68, 850–858
30. Leptos, K. C., Sarracino, D. A., Jaffe, J. D., Krastins, B., and Church, G. M.
(2006) MapQuant: open-source software for large-scale protein quanti-
fication. Proteomics 6, 1770–1782
monoisotopic masses and ion populations for large biomolecules from
32. Olsen, J. V., de Godoy, L. M. F., Li, G., Macek, B., Mortensen, P., Pesch, R.,
Makarov, A., Lange, O., Horning, S., and Mann, M. (2005) Parts per
million mass accuracy on an Orbitrap mass spectrometer via lock mass
effects in Fourier transform mass spectrometry. II. Mass calibration.
Anal. Chem. 56, 2744–2748
for Fourier transform-ion cyclotron resonance/mass spectrometry. Spectro-
tos. World 2, 24–30
13, 99–106
and external mass calibration laws utilized in Fourier transform ion
cyclotron resonance mass spectrometry. Anal. Chem. 77, 2406–2414
chemical formula identification of aromatic hydrocarbons and aromatic
compounds containing sulfur, nitrogen, or oxygen in petroleum distillates
and refinery streams. Anal. Chem. 68, 46–71
evaluation of mass spectrometry platforms used in large-scale proteome
investigations. Nat. Methods 2, 667–675
Jonscher, K. R., Pierce, K. G., Old, W. M., Cheung, H. T., Russell, S.,
Improving reproducibility and sensitivity in identifying human proteins by
shooter proteomics. Anal. Chem. 76, 3556–3568
ion trap mass spectrometer. J. Am. Soc. Mass Spectrom. 9, 1175–1188
immobilized metal (Fe3+) affinity chromatography. Anal. Biochem. 154,
250–258
42. Piccirillo, S., McLeod, M. L., Stukkenberg, P. T., Burke, D. J., Ross,
proteome analysis by mass spectrometry and its application to Saccha-
romyces cerevisiae. Nat. Biotechnol. 20, 301–305
performance technique of mass analysis. Nat. Methods 7, 1156–1162
Spectrom. 36, 1083–1091
Chem. 78, 493–500
versus low mass accuracy MS/MS for complex mixture analysis in pro-
teomics, in Proceedings of the 53rd ASMS Conference on Mass Spec-
trometry, San Antonio, TX, June 5–9, 2005
directional synonymous codon usage bias, and its potential applica-
tions. Nucleic Acids Res. 15, 1281–1295
48. Ghazemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A.,