research articles Journal of proteome research

Quality Control Metrics for LC−MS Feature Detection Tools Demonstrated on *Saccharomyces cerevisiae* Proteomic Profiles

Brian D. Piening,[†] Pei Wang,[†] Chaitanya S. Bangur,[†] Jeffrey Whiteaker,[†] Heidi Zhang,[†] Li-Chia Feng,[†] John F. Keane,[†] Jimmy K. Eng,[†] Hua Tang,[†] Amol Prakash,[†,‡] Martin W. McIntosh,[†] and Amanda Paulovich*[,†]

Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N, Seattle, Washington, 98109 and Department of Computer Science and Engineering, University of Washington, Seattle, Washington, 98195

Received December 2, 2005

Quantitative proteomic profiling using liquid chromatography−mass spectrometry is emerging as an important tool for biomarker discovery, prompting development of algorithms for high-throughput peptide feature detection in complex samples. However, neither annotated standard data sets nor quality control metrics currently exist for assessing the validity of feature detection algorithms. We propose a quality control metric, *Mass Deviance*, for assessing the accuracy of feature detection tools. Because the Mass Deviance metric is derived from the natural distribution of peptide masses, it is machine- and proteome-independent and enables assessment of feature detection tools in the absence of completely annotated data sets. We validate the use of *Mass Deviance* with a second, independent metric that is based on isotopic distributions, demonstrating that we can use *Mass Deviance* to identify aberrant features with high accuracy. We then demonstrate the use of independent metrics in tandem as a robust way to evaluate the performance of peptide feature detection algorithms. This work is done on complex LC−MS profiles of *Saccharomyces cerevisiae* which present a significant challenge to peptide feature detection algorithms.

Keywords: yeast proteomics • bioinformatics • *S. cerevisiae* • feature detection

Introduction

Quantitative profiling by liquid chromatography−mass spectrometry (LC−MS) is rapidly becoming a common and powerful technique in proteomics that provides important insight into cellular biology and medicine.[1] Currently, there are two main approaches using LC−MS for peptide quantification from complex samples. The first approach involves the incorporation of an isotopic label followed by quantification of peptide features by LC−MS and subsequent identification by shotgun tandem mass spectrometry (MS/MS).[2,3] Though current hybrid mass spectrometers, such as the LTQ-FT, can acquire LC−MS and LC−MS/MS information simultaneously in a single experiment, undersampling of ions by LC−MS/MS remains a significant challenge as a large number of quantifiable peptides go unidentified in a single analysis.[4] An alternative label-free approach has been employed recently that involves feature detection, quantification, and alignment of a large number of peptide "fingerprints" across a series of biological replicates to produce a feature table that is similar to that produced by a microarray experiment.[5,6] This is followed by common statistical approaches (clustering, marker analysis) to search for interesting biological relationships between two or more classes of biological samples. Because discovery and quantification occur at the peptide fingerprint level, only statistically relevant peptides are selected for identification either via targeted LC−MS/MS or by using an accurate mass and time tag database.[7] To facilitate the peptide fingerprinting strategy, a number of high-throughput peptide feature detection methods have recently been developed based on a variety of techniques.[8−11] However, there has yet to be significant progress in determining what percentage of features identified by these algorithms correspond to actual peptides. To assess the false discovery rate of peptide feature detection methods, there needs to be some way of measuring absolute truth in LC−MS samples. Such a data set could be attempted with a mixture of synthetic peptides, but would not provide adequate complexity needed to completely evaluate algorithms normally used for biomarker discovery on complex tissue and fluid proteomic samples. Using shotgun MS/MS data to generate such a data set on complex samples is not ideal because of the aforementioned undersampling problem, as a large number of LC−MS peaks in the complex sample will go unevaluated by collision-induced dissociation (CID). This undersampling would greatly affect the low intensity features in a complex sample, which exist near the limit of detection of the instrument. These low intensity features present a significant challenge for feature detection algorithms and thus cannot be excluded from any assessment of quality.

In this manuscript, we describe a method that allows investigators to evaluate peptide feature detection on highly complex samples, but does not require completely annotated data sets. We introduce a metric, the *Mass Deviance*, which compares detected features to theoretical peptide mass distri-

* To whom correspondence should be addressed. E-mail: apaulovi@fhcrc.org.
[†] Fred Hutchinson Cancer Research Center.
[‡] University of Washington.

10.1021/pr050436j CCC: $33.50

Published on Web 05/27/2006

butions to assess feature quality. This method is based on the fact that peptides can only consist of a discrete subset of masses.[12] Thus, if we examine any particular 1 Da mass range, there are portions of that 1 Da window where peptides are concentrated and other areas where peptides do not exist.[13] These areas without any theoretical peptides occur at different decimal ranges dependent on the integer mass value in question. Because mass spectra are recorded in terms of mass-to-charge ratios and not mass, peptides can assume additional values within the 1 Da mass-to-charge windows if they are of higher order charge states. As we will show, this phenomenon can be easily seen in complex LC−MS profiles, such as that of the model organism *Saccharomyces cerevisiae*. Our metric maps peptide mass fingerprints identified by feature detection algorithms against mass distributions based on theoretical tryptic digests of protein databases. If a peptide occurs in an area of low to zero probability, then it is flagged as a possible example of an error made by the feature detection algorithm.

Because the metric is based on naturally existing peptide mass distributions, it is LC−MS platform-independent and organism-independent. We apply this metric to peptide feature detection in complex yeast LC−MS profiles to characterize the performance of a recently developed feature detection algorithm. To validate the use of this method, we introduce another metric, the *Isotope Deviance*, as a way to corroborate results obtained with the *Mass Deviance* metric. The *Isotope Deviance* assesses the isotopic distribution of a peptide fingerprint in comparison to an empirical model. We observed a 95% correlation between *Mass Deviance* and *Isotope Deviance* in flagging incorrect peptide features.

Materials and Methods

Yeast Protein Extraction. *S. cerevisiae* strain BY4741 (MATa, leu2D0, met15D0, ura3D0, his3D1) was grown in synthetic complete medium at 30 °C to log phase, harvested and diluted into 30 mL synthetic complete media to a cell count of 3×10^6/mL. After 4 h of growth at 30 °C cells were harvested and washed three times with ice cold dH_2O. Cells were lysed by incubation with 1 mL of ice cold 10% TCA for 1 h at 4 °C. Protein precipitates were collected by centrifugation, washed twice with 1 mL cold 90% acetone, and dried in a SpeedVac (Thermo Savant, Holbrook, NY). Proteins were solubilized in 300 μL of 8 M urea, 50 mM ammonium bicarbonate and reduced by incubation at 56 °C for 1 h in the presence of 15 mM DTT.

Yeast Protein Digestion. A 150-μL portion of reduced yeast protein lysate was alkylated in 25 mM of iodoacetamide and diluted to 1 M urea with 50 mM ammonium bicarbonate. 12 μg of Trypsin Gold (Promega, Madison, WI) was added to the reduced and alkylated yeast protein lysate and trypsin digest was carried out overnight at 37 °C. Trypsin was inactivated by addition of glacial acetic acid and insoluble material was removed by centrifugation. Peptides were purified using 3 mL SPEC C18 columns and concentrated in a SpeedVac. Peptide concentration was estimated by absorbance at 280 nm, using a reference curve generated using synthetic tripeptides.

LC−TOF MS Profiling. Each independent peptide digest was analyzed by nanoLC−MS. The HPLC system (Agilent 1100 nanoflow system) was configured with a solvent degasser, microautosampler, and isocratic pump (for sample loading). Briefly, 1.4 μg (∼28 pmol) of peptide digest were loaded onto a C_{18} precolumn (100 μm × 1.5 cm, Integafrit, New Objective, Woburn, MA; 5 μm Atlantis packing, Waters, Milford, MA) with 2% acetonitrile, 0.1% formic acid at a flow rate of 20 μL/min and washed for 5 min. Peptides were eluted with a gradient of 2−40% solvent B (acetonitrile, 0.1% formic acid) over 120 min on a capillary C_{18} silica-based monolithic column (100 μm × 15 cm, Chromolith CapRod, Merck, Darmstadt, Germany) at a flow rate of 800 nL/min. The column was connected to a capillary fused silica emitter (90 μm O. D. × 20 μm I. D. × 5 cm length) via a true zero dead volume connector (Upchurch Scientific, Oak Harbor, WA). LC−MS profiles were obtained on an LCT Premier time-of-flight mass spectrometer (TOF-MS, Waters) equipped with a nanolock spray electrospray source and operated over m/z range 400−1600 for 1.0 s with a 0.05 s interscan delay time. The general mass spectrometric instrumental parameters were as follows: capillary voltage, 3 kV; cone voltage, 60 V; source temperature, 120 °C. Mass calibration was performed using a sodium formate calibration mix, and a lock-mass reference standard (glu-fibrinopeptide, 300 fmol/μL) was infused in the reference spray at 1.0 μL/min for accurate mass determination.

Linear Ion Trap Tandem MS. Samples were analyzed by nanoLC−MS/MS using an nanoflow HPLC system (Agilent 1100, Agilent) connected to a linear ion trap mass spectrometer (LTQ, Thermo Electron). Yeast cell lysate (1.4 ug, ∼28 pmol) was loaded onto a C_{18} precolumn (100 μm × 1.5 cm, Integafrit, New Objective, Woburn, MA; 5 μm Atlantis packing, Waters, Milford, MA) with 2% acetonitrile, 0.1% formic acid at a flow rate of 20 μL/min and washed for 5 min. Peptides were eluted with a gradient of 2−40% solvent B (acetonitrile, 0.1% formic acid) over 120 min on a capillary C_{18} silica-based monolithic column (100 μm × 15 cm, Chromolith CapRod, Merck, Darmstadt, Germany) at a flow rate of 800 nL/min. The column was connected to a capillary fused silica emitter (90 mm O. D. × 20 mm I. D. × 5 cm length) via a zero dead volume connector (Upchurch Scientific, Oak Harbor, WA). Typical LTQ instrument settings include a spray voltage of 1.5 kV, an ion transfer tube temperature of 200 °C, and a collision gas pressure of 1.3 Torr. Voltages across the capillary and the quadrupole lenses were tuned for optimal signal intensity using the +2 ion of angiotensin I (m/z 649). Blank runs, where only buffer was injected, were also performed using the same LC−MS/MS methods.

FTICR, Linear Ion Trap Hybrid MS and MS/MS. LC−ESI−MS/MS experiments were performed on an Eksigent (Dublin, CA) nanoLC-2D HPLC coupled with a ThermoElectron (Waltham, MA) LTQ-FT mass spectrometer. The instruments were configured as in Yi et al.,[14] utilizing New Objective (Woburn, MA) 75 um IntegraFrit and PicoTip products for the trapping and analytical columns, respectively. A solvent system of 0.1% formic acid (A) with 0.1% formic acid in acetonitrile (B) was used at a flow rate of 300 nL/min. Data were collected in a data-dependent mode in which a high mass resolution/high mass accuracy profile scan in FT was followed by centroided MS/MS scans of the five most abundant ions from the preceding MS scan in the ion trap. The isolation width for precursor masses was set to ±1.5 Da and the normalized collision energy was set to 30%. The five selected ions for tandem MS were placed on an exclusion list for 3 min.

Informatics. LC−MS profiles from the electrospray TOF mass spectrometer were converted from RAW format to mzXML using *massWolf*.[15] Peptide feature detection from LC−MS profiles was performed using the *msInspect* feature detection tool.[9] LC−MS/MS data were searched using the COMET search algorithm.[16] Peptides and proteins were assigned probabilities using PeptideProphet[17] and ProteinProphet[18] tools,

respectively. The protein sequence database used in this study for MS/MS searches was the November 4, 2005 release of the *Saccharomyces* Genome Database (ftp://ftp.yeastgenome.org/yeast/) for yeast, and the December 9, 2004 IPI database from EMBL−EBI (http://www.ebi.ac.uk/IPI/). Theoretical peptide digests for comparisons were computed from the SGD sequence database using in-house software written in Java. Comparisons from these digests to detected peptide features were done using R. All software written for this study is available upon request.

Quality Control. We introduce the following quantities to investigate various peptide properties:

(1) *Mass Decimal* of a peptide is the decimal fraction after the integer value in the molecular weight. This quantity helps to illustrate the degree of discreteness of the distribution of mass values of yeast peptides.

(2) *Mass Deviance* of an observed mass value (M_{obs}) is defined to be the distance from M_{obs} to the nearest theoretical tryptic peptide ($\{m^i_{\text{the}}\}_i$):

$$(1)\ \mathit{MassDeviance}(M_{\text{obs}},\{m^i_{the}\}_i) = \min_i |M_{\text{obs}} - m_{\text{the}}{}^i|.$$

This value can serve as a quality measurement for peptide charge estimation in feature detection. The more discrete the distribution of $\{m^i_{the}\}_i$ is, the greater the power *Mass Deviance* will have for discriminating correct features from incorrect features. We demonstrate in the results section that this measurement provides a simple and efficient way to identify incorrect peak calls in feature detection.

(3) *Isotope Deviance* of one feature is defined as the L2 norm distance between the observed isotopic peak distribution for a peptide and the "ideal" distribution for the same mass and charge value (minus experimental noise and variation). We can approximate the "ideal" isotopic shape empirically for peptides of a certain charge and mass by averaging all feature spectra of the same charge state and similar mass values. This is because at a given mass, the variation of isotopic shapes resulting from the differences between sequences of amino acids is usually much less than the variation introduced by experimental noise; thus we assume that the peptides with similar mass values and at the same charge states would have similar isotopic patterns. We also generated a model using the "averagine" method, which utilizes average amino acid masses to generate a composite spectrum at discrete mass values.[19] This method would be useful for characterizing data sets of lower complexity, but does not take into account instrument-specific resolution and noise effects. For our yeast sample, we found the empirical model to have better discriminating power. A comparison of the two methods can be found in the Supporting Information.

We denote the mass value, charge state, and the observed spectrum vector of the ith feature as M_i, Ch_i, and $X_i = (x_i^1,\ldots,x_i^{217})$. The corresponding empirical isotope shape vector $Y_i = (y_i^1,\ldots,y_i^{217})$ is calculated as

$$(2)\ y_i^j(M_i,Ch_i) = \underset{k}{\text{ave}}\{x_k^j/||X_k||_{L1}: Ch_k = Ch_i, |M_k - M_i| < \delta\}$$

where δ is a small constant. We normalize the spectrum vectors via the L1 norm to form an (intensity-independent) estimate of the isotopic shape distribution.

For each feature, we use a vector to represent its average mass spectra from the three contiguous scans with the highest peak intensity. This vector records the intensity measurements at 217 m/z points, or pixels, evenly spaced in the $[-0.5, 3]$ m/z neighborhood interval of the feature's monoisotopic position. Since a peptide's isotopic distribution provides key information for estimating m/z and charge values, these vectors can be used in turn to assess the feature detection quality.

We then define:

$$(3)\ d(X_i,Y_i) = ||\frac{X_i}{||X_i||_{L1}} - \frac{Y_i}{||Y_i||_{L1}}||_{L2},$$

which measures the L2 norm of the distance between the normalized observed spectrum vectors and the corresponding empirical isotopic shape. If the observed isotopic shape of a peptide is similar to the ideal shape, i.e., then the (M_i,Ch_i) estimation is correct, we would expect to get a small d value. Thus, we can use this d value to confirm peptide features as incorrect based on the isotopic distribution.

We can also test whether(M_i,Ch_i) estimation is correct by testing whether the d value significantly deviates from 0. If (M_i,Ch_i) is correct, then we can assume the following

$$(4)\ x_i^j = y_i^j + \epsilon_i^j,\ j = 1,2,\ldots,127, \epsilon_i^j \overset{i.i.d.}{\sim} N(0,\sigma^2).$$

Thus, d/σ would follow a standard Chi-square distribution with degrees of freedom equal to 127.

So the p value for the (M_i,Ch_i) estimation being correct can be calculated as follows

$$(5)\ P(d) = \Pr\left(\chi^2_{127} > \frac{d}{\hat{\sigma}}\right),$$

where $\chi_{127}{}^2$ denotes the random variable following the distribution of Chi-square(df = 127), and

$$\hat{\sigma} = \frac{\text{median}(d)}{127}$$

(here we assume that the majority of estimations of mass and charge are correct).

Results and Discussion

Unfractionated yeast LC−MS proteomic profiles represent a significant challenge for peptide feature detection algorithms. Though it may seem paradoxical because of the small size of the yeast proteome (~6000 proteins), LC−MS profiles of yeast are very complex, much more so than that of mammalian plasma, for example.[20] This is due to the smaller range of yeast protein abundances which results in large numbers of peptide identifications in a single LC−MS analysis. Using the peak detection capabilities of *msInspect*, a feature detection tool developed in-house, we observed a total of 24 886 peptide features (deisotoped but not deconvoluted) in a single yeast LC−MS sample. Furthermore, cumulative peptide identifications from 31 LC−MS/MS repeat analyses of the yeast cell lysate resulted in the identification of 4550 unique peptides meeting a PeptideProphet score of 0.95 or higher (Figure 1). Although the nearly 25 000 yeast peptide features profiled on the TOF far exceed the number of peptides identified by MS/MS, the TOF features represent peptides present in multiple charge states, peptides that do not fragment well by collision-induced dissociation, modified or polymorphic peptides not present in the database, and low-confidence features misidentified by feature detection algorithms.

At the protein level, we observed 1279 unique yeast proteins at a 5% error rate cutoff (ProteinProphet ≥ 0.70). This covers

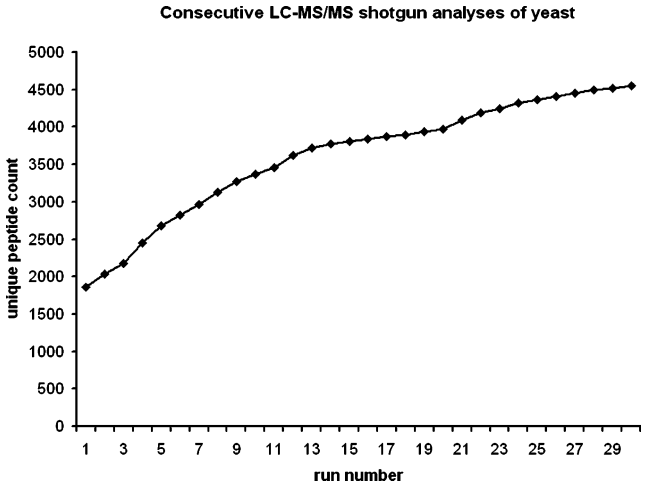

Figure 1. Cumulative number of peptide identifications over 31 LC−MS/MS repeat analyses of unfractionated yeast lysate (PeptideProphet ≥ 0.95).

22.0% (1279/5823) of the protein sequences in the database. Comparing these identifications to published cellular abundance data from the O'Shea and Weissman labs,[21] we observe proteins sequenced by MS/MS down to ~50 molecules/cell, representing 5 orders of magnitude of protein abundance. Figure 2 shows protein coverage by MS/MS vs cellular abundance, in which we observe in general that we achieve high sequence coverage by MS/MS for high abundance proteins, as expected.[21]

Because of their complexity, yeast LC−MS profiles present a particularly useful platform for illustrating the natural distribution of peptide masses on which our *Mass Deviance* metric is based. From these dense profiles, we can observe a clear periodicity of peptide mass distributions in the LC−MS data. For example, in Figure 3, we show images of the yeast LC−MS profile in the small *m*/*z* regions 500−505 Da and 1000−1005 Da, where one can clearly observe peptides arranged into areas of high signal intensity at regular mass/charge intervals across the time domain. This can be seen as the horizontal bands of high intensity in Figure 3A and 3B, and consist of dense regions of peptides with regularly spaced mass values, overlapping in chromatographic time. We observed that the intervals became more diffuse at higher *m*/*z* values, as seen in Figure 3B. To confirm that the banding pattern is not machine-specific, we examined the same yeast sample using a LTQ-FT, and also observed the same patterns in the resultant data. (Figures for the LTQ-FT data are available on the Supporting Information website.)

To confirm that these bands are not due to baseline noise, we ran a series of "blanks," 10 min gradients injecting only

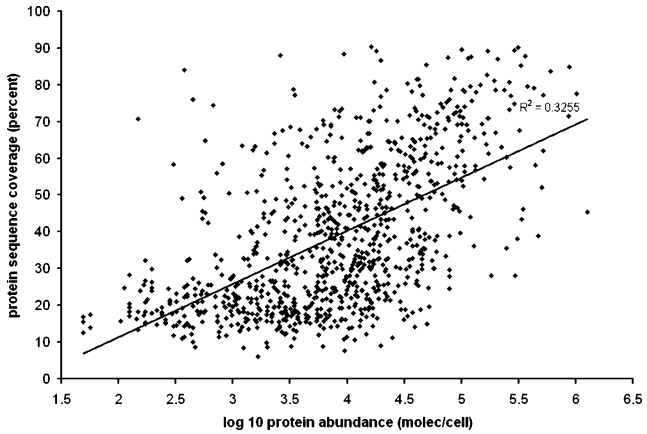

Figure 2. Protein coverage from shotgun LC−MS/MS data vs cellular abundance (molecules/cell). Abundance data are as reported by Ghaemmaghami el al.[21]

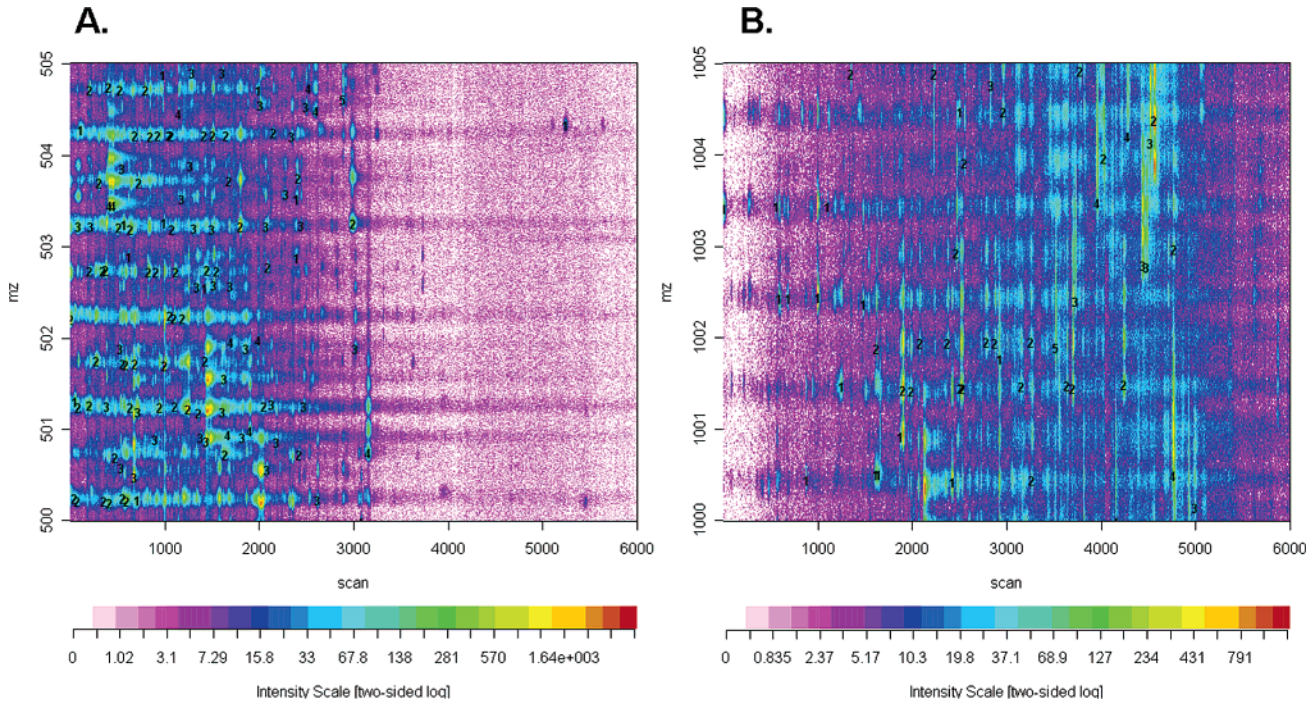

Figure 3. LC−MS profile image of yeast lysate in the *m*/*z* region A [500, 505] and B [1000,1005]. This multidimensional image indicates scan number (*x*-axis), *m*/*z* value (*y*-axis), and signal intensity plotted as a color scale. Peptide features identified by *msInspect* are indicated by their numbered charge state.

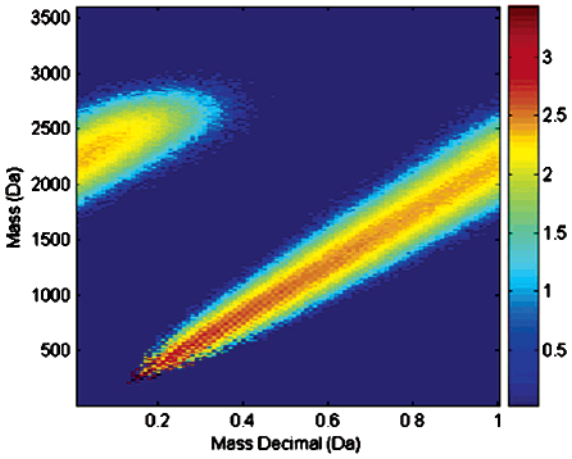

Figure 4. Deltamass distribution of theoretical peptide digest. The peptide mass is plotted vs the delta mass (decimal) as a heatmap. The color scale represents the log number of peptides at a given location. At specific mass values, peptides only have specific mass defects.

buffer prior to each lysate run (data not shown). Although we did not detect regularity in the baseline of a single scan, we did notice subtle 1 Da periodicity after summing over many scans (~100). This noise is common in electrospray data,[22,23] but its signal intensity is too low to be the sole cause of our 1 Da bands.

The interval distribution of peptides that we have described can be attributed to the fact that there are a limited number of combinations of masses of amino acid residues that a peptide can assume. The range of decimal values also becomes more diffuse at higher masses, due to the growing number of possible combinations of amino acids.[13] To illustrate this phenomenon, we computed an in silico tryptic digest of the yeast proteome database, generating over 300 000 tryptic peptide sequences. For each peptide, we computed the *Mass Decimal* (the decimal value after the integer, described in the *Methods* section), which was plotted vs molecular weight in Figure 4. From the figure, we observe that the decimal masses of peptides may only assume values in a distinct range, which shifts with increasing mass.

The periodicity of peptide masses provides a useful metric for assessing the accuracy of feature detection algorithms. For example, features that fall outside of the natural distribution must represent errors made by the feature detection tool. To test this prediction, we sought to compare the theoretical digest of the yeast protein database to features empirically identified with the feature detection tool, *msInspect*. Figure 5 details the theoretical distribution of peptide m/z values in the mass ranges [500, 550] and [1000, 1050] vs the observed distribution of peptides detected by *msInspect* for the same ranges. The integer value was stripped off leaving only the decimal, which was plotted vs frequency of occurrence for different charge states. From this figure, it can be clearly seen that peptide decimal m/z values fall into regular bands with no peptides in between. It can also be observed that the theoretical distributions differ for higher charge states, which is due to the fact that a higher order charge state represents a higher peptide mass, which takes on a different probability distribution. In other words, while the +2 distribution within [500 550] will be derived from peptides of mass 1000−1100 Da, the +3 distribution will be derived from peptides of mass 1500−1650. Our observed distributions of features, detected by *msInspect* and shown in the right panel of Figure 5, closely resemble the theoretical distributions, thus explaining the observed m/z bands in our LC−MS data. However, there are areas of the observed distribution that deviate from the theoretical. This can be seen in the observed distribution for charge +2 features,

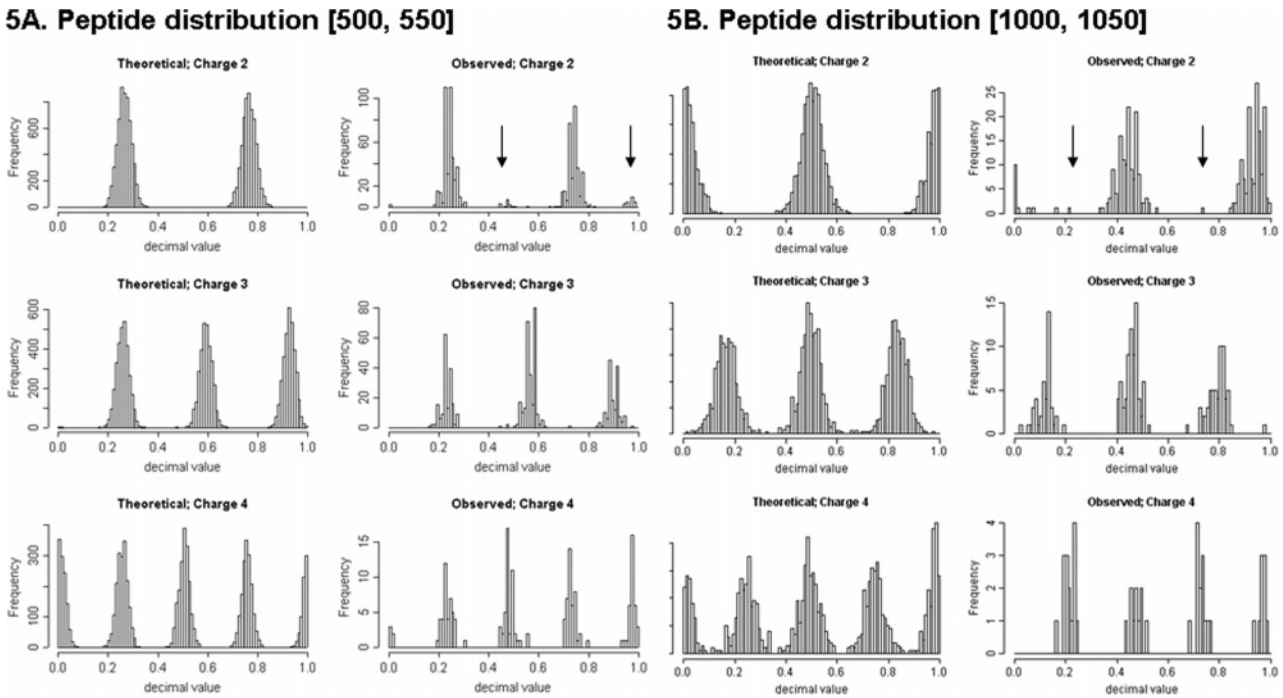

Figure 5. (A) Left three panels show the distribution of decimal values of mz (mass defects) for those theoretical peptides with charge 2, 3, or 4 and m/z in [500, 550], whereas the right three panels show the same distribution for observed peptides. (B) Distributions over the m/z interval [1000, 1050]. Arrows indicate areas where the experimental data deviate from theoretical.

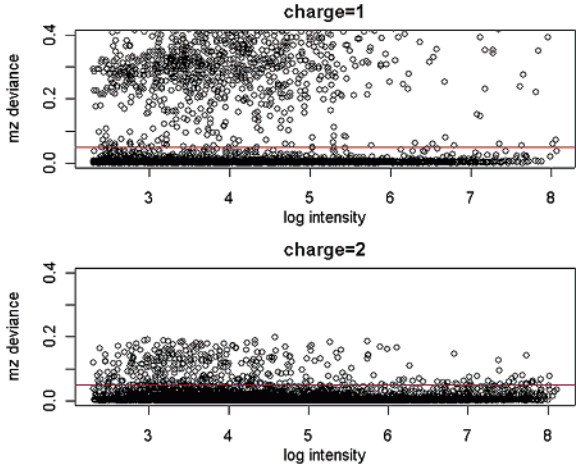

Figure 6. Difference in mass-to-charge value from nearest theoretical value, which is designated the *m*/*z* deviance, is plotted vs the peptide intensity for all features detected from the yeast LC−MS profiles.

which contains two smaller peaks that are not present in the theoretical data, indicated by arrows in the figure.

On the basis of this observation, we introduce a quantity, the *Mass Deviance*, which can be used to evaluate feature detection algorithms based on the similarity of mass values of identified features to mass distributions generated from theoretical peptide digests *(see Quality Control in the Methods section)*. Since the theoretical mass values of peptides concentrate tightly on discrete bins along the mass axis, features that clearly deviate in *m*/*z* from these theoretical bins most likely result from misassignment of the charge state or monoisotopic peak by the feature detection tool. Thus, features with a high *Mass Deviance* are flagged and referred to as low-confidence features below.

Among the total features detected by *msInspect* in our yeast profile, 8.36% are low-confidence features. Perhaps not surprisingly, these low-confidence features are biased toward lower intensities, where feature detection tools would be most challenged (Figure 6). A large percentage of charge +1 features are in the low-confidence category, which is partly due to the overall low intensities of charge +1 features, and also due to the fact that quality assessment with theoretical mz values is most effective for charge +1 features. This effectiveness is because the theoretical *m*/*z* regions are generally more discrete for +1 peptides than for higher charge states (Figure 5) and occupy less of the spectrometer's total *m*/*z* range.

To assess the specificity of the *Mass Deviance* metric, we used data generated on the LTQ-FT of the yeast lysate to compare precursor peaks identified by msInspect with their corresponding sequences by MS/MS. Of the ~1300 peptides identified by MS/MS, we calculated the *Mass Deviance* on the corresponding precursor mass as identified by msInspect. Features with high scores of sequence identification (PeptideProphet ≥ 0.95) were defined as correct peptides. For different cutoffs of *Mass Deviance*, the false discovery rate (FDR) is calculated as the percent of correct peptides among the features assigned to the low-confidence category (with *Mass Deviance* greater than the cutoff). The result is illustrated in Figure 7. At a stringent *Mass Deviance* cutoff of 0.02, less than 4% of the low-confidence features are correct peptides, demonstrating the *specificity* of the metric for accurately identifying true peptide features.

It is impossible to ascertain the true *sensitivity* of the metric

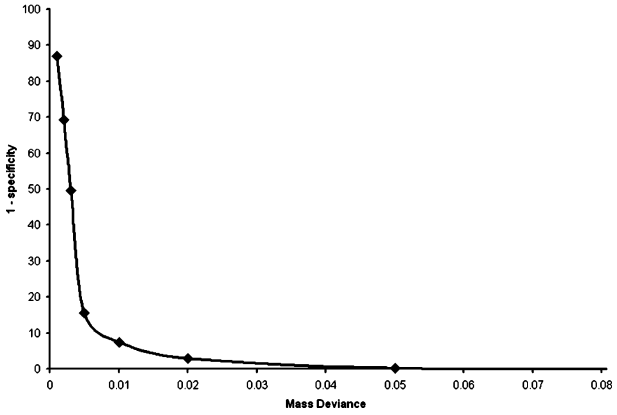

Figure 7. FDR of the *Mass Deviance* metric. The percentage of features incorrectly flagged is plotted for different *Mass Deviance* cutoffs (a lower deviance value is more stringent). An incorrect call is defined as a feature that was flagged by *Mass Deviance* but later verified by MS/MS identification.

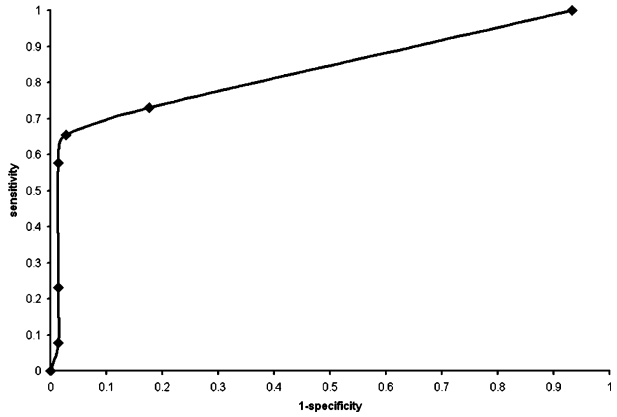

Figure 8. ROC plot showing the sensitivity and specificity of the Mass Deviance metric for 100 randomly chosen features that were hand annotated to determine true feature quality.

because a large number of features identified by the feature detection algorithm do not have a valid MS/MS identification associated with them. This is primarily due to peptides with poor fragmentation and the undersampling of ions by MS/MS. For a sample of this complexity, it is not possible to know the identity of every peptide present. We attempt to estimate the sensitivity by choosing at random 100 features of varying Mass Deviance. These features were then hand-curated to determine whether they represented true peptides, and the ROC plot detailing the sensitivity and specificity is shown in Figure 8. From the hand curation of the 100 random features, 26 were determined to be poor features, which consisted of a mixture of incorrect charge state identifications, wrong monoisotopic mass identifications (usually associated with overlapping isotope distributions of coeluting peptides), and selection by the feature detection tool of portions of the baseline as features. At the same cutoff of *Mass Deviance*, the specificity estimation based on the hand-curated feature set is similar (within 15%) to the above result using MS/MS identifications, despite the small sample size. According to the ROC in Figure 8, the optimal cutoff value for *Mass* Deviance is chosen to be 0.05, which corresponds to a sensitivity of 65% and a specificity of 97%.

Additionally, we developed a second metric we call *Isotope Deviance* (See Quality Control in the Methods section) to use

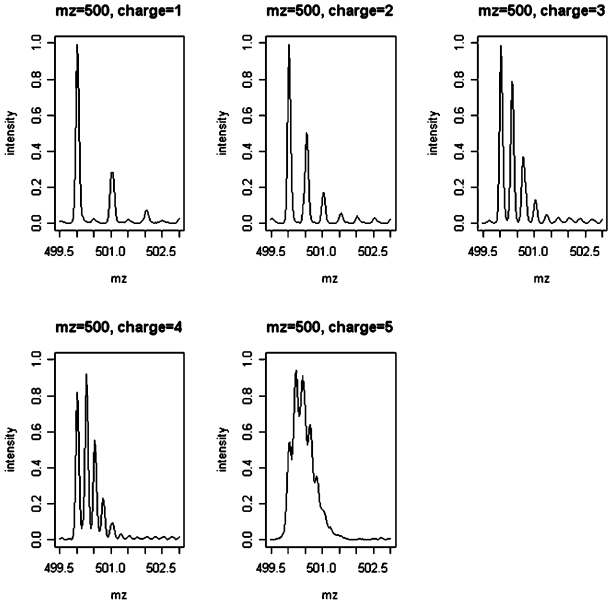

Figure 9. Isotopic distribution for peptides with $m/z = 500$ and different charge states generated from the sum of all peptides detected.

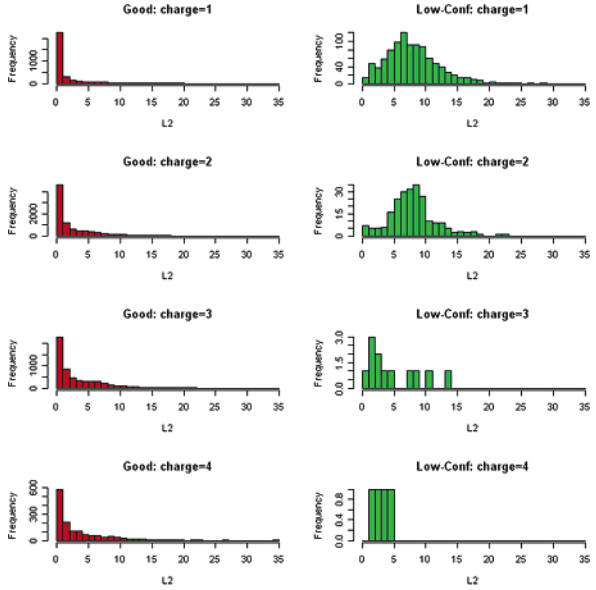

Figure 10. Correlation between the Mass Deviance and Isotope Deviance scores. In the left panel, we plot the frequency of *Isotope Deviance* scores (*x* axis) by charge state among features detected by *msInspect* and determined to be of "high-confidence" using the *Mass Deviance* metric. The right panel shows the same histogram for features that did not pass the *Mass Deviance* filter, and are thus flagged as possible errors made by the feature detection tool.

as an independent means of corroborating the accuracy of the Mass Deviance metric in detecting low confidence features. This metric compares individual isotopic ratios to a model distribution as follows: On the basis of the average of isotopic ratios of all features, we generated a model isotopic distribution at each m/z value. An example is seen in Figure 9, where for each charge state a model distribution was generated at $m/z = 500$. Then for each feature, we compared its isotope shape with the empirical template. This was done by normalizing each feature and calculating the L2 norm distance between the observed isotopic pattern and empirical pattern. This score is defined as the *Isotope Deviance*. If using Mass Deviance accurately identifies misassigned features, we would expect a high concordance between *Mass Deviance* and *Isotope Deviance*.

To assess the concordance between *Mass Deviance* and *Isotope Deviance*, we used both methods in parallel to assess the peptide features detected in our LC−MS data set by msInspect. We computed the *Isotope* Deviance for each peptide feature and plotted the distribution of scores for features classified as high- and low-confidence by the *Mass Deviance* metric (Figure 10). Comparing the distributions of high- and low-confidence features, we observe that the low-confidence features in the +1 and +2 charge states have much higher *Isotope Deviance* scores than the high-confidence features, suggesting that the *Mass Deviance* metric successfully identified aberrant features. Note that at +3 and higher charge states the metric is somewhat less effective because the mass defect becomes more diffuse for the high mass values that typically accompany high charge states for peptides in the range of the detector. However, the majority of peptides typically seen in LC−MS data fall into the +2 or below category.

Finally, we computed the correlation between low-confidence features identified by the *Mass Deviance* and *Isotope Deviance* metrics. Results of this assessment for different charge states can be seen in Table 1. Among the low-confidence features identified by the *Mass Deviance* metric, 87% are also identified as incorrect feature detections based on the *Isotope Deviance* ($p < 0.05$). This high degree of correlation between two different methodologies for assessing feature quality suggests that both metrics identify low-confidence features with high specificity. By combining these metrics, we identified a total of 1167 features that possessed both a high *Mass* Deviance and high *Isotope Deviance*, indicating that these features have both aberrant mass values and atypical isotopic distributions. Because these features were flagged independently by two distinctly different quality control metrics, they likely result from an incorrect charge state or incorrect monoisotopic mass assignment, and could be manually corrected or filtered out. Alternatively, these features may be due to chemical noise or other molecules that do not resemble peptides, and could be identified through other techniques such as NMR if desired.

Table 1. Relationship between *Mass Deviance* and *Isotope Deviance* Assignments for Different Charge States[a]

isotope deviance	charge 1	charge 2	charge 3	charge 4	charge 5
low-confidence	929	225	8	3	2
high-confidence	55	101	4	1	2

[a] From the 1330 features identified as low-confidence by *Mass Deviance*, we classify each based on the *Isotope Deviance* metric. Features in the low-confidence row were confirmed as aberrant features by the *Isotope Deviance* filter at $P < 0.05$.

Conclusions

Peptide fingerprinting shows significant potential for quantitative proteomics when coupled to downstream MS/MS. However, stringent quality control and standards for comparison of both tools and LC−MS data across laboratories and instruments is challenging due to a lack of QC metrics and a lack of annotated data sets. Because peptide sequence identification occurs in a separate step, a high quality set of peptide fingerprints needs to be selected for statistical analysis, and

much emphasis has to be placed on quality feature detection tools. Because of the dynamic range of most complex samples, the majority of peptides in LC−MS profiles occur at low intensity, presenting an intrinsic challenge for identification by feature detection algorithms. Incorrect peptide fingerprints based on misidentified peaks in LC−MS data will lead to a loss of statistical power in downstream analysis, and hence interesting biological phenomena may be missed.

The high-throughput sequencing that ushered in the genomic revolution was made possible by the PHRED[24] metric, and thus similar QC metrics need to be developed for comparative proteomics. Quality assessment has been implemented at the MS/MS database search level,[17,18] but not at the MS1 fingerprint level. Because of undersampling of ions and poor fragmentation of some peptides, it is impossible to know the identity of every peak in complex LC−MS data, making any measure of absolute truth difficult and thus creating an obstacle to validating feature detection algorithms.

We describe two platform- and proteome-independent quality control metrics, the *Mass Deviance* and *Isotope Deviance,* for evaluating peptide feature detection, and demonstrate their utility on high-complexity LC−MS fingerprints of yeast. Because these metrics are derived from natural peptide characteristics, they are not dependent on highly annotated data sets for assessing the accuracy of feature detection tools. The *Mass Deviance* metric will be more beneficial to instruments with high mass accuracy, such as an FTICR, because the high mass accuracy will enable more efficient identification of peptides with deviant mass values. The *Isotope Deviance* will be best utilized on high resolution instrumentation, as the metric is based on comparing isotopic distributions of peptides to an empirical model. These methods could be combined with other models based on different peptide properties as well. We propose that these metrics can ultimately be combined with additional metrics for assessing accuracy of feature quantification as well as feature alignment across runs, providing quality assessment in all stages of the peptide fingerprinting approach.

Acknowledgment. This work was funded in part by NCI contract SAIC, NCI−Frederick 23XS144A, Biomarker Discovery Initiative. The authors would like to thank Jenny Chen and Phil Gafken for FTICR data and Richard Ivey for reviews and suggestions.

Supporting Information Available: A comparison of the two methods presented. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Ong, S. E.; Mann, M. *Nat. Chem. Biol.* **2005**, *1*, 252−262.
(2) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turacek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17*, 994−999.
(3) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. *Mol. Cell. Proteomics* **2002**, *1*, 376−386.
(4) Masselon, C.; Pasa-Tolic, L.; Tolic, N.; Anderson, G. A.; Bogdanov, B.; Vilkov, A. N.; Shen, Y.; Zhao, R.; Qian, W. J.; Lipton, M. S.; Camp, D. G. 2nd; Smith, R. D. *Anal. Chem.* **2005**, *77*, 400−406.
(5) Wang, W.; Zhou, H.; Lin, H.; Roy, S. Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H. *Anal. Chem.* **2003**, *75*, 4818−4826.
(6) Wiener, M. C.; Sachs, J. R.; Deyanova, E. G.; Yates, N. A. *Anal. Chem.* **2004**, *76*, 6085−6096.
(7) Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R. *Proteomics* **2002**, *2*, 513−523.
(8) Li, X. J.; Yi, E. C.; Kemp, C. J.; Zhang, H.; Aebersold, R. *Mol. Cell. Proteomics* **2005**, *4*, 1328−1340.
(9) Bellew, M.; Coram, M.; Igra, M.; Fitzgibbon, M.; Randolph, T.; Wang, P.; Eng, J.; Lin, C.; Goodlett, D.; Fang, R.; Detter, A.; Zhang, H.; Whiteaker, J.; Paulovich, A.; McIntosh, M., in preparation.
(10) Radulovic, D.; Jelveh, S.; Ryu, S.; Hamilton, T. G.; Foss, E.; Mao, Y.; Emili, A. *Mol. Cell. Proteomics* **2004**, *3*, 984−997.
(11) Hastings, C. A.; Norton, S. M.; Roy, S. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 462−467.
(12) Mann, M. *Abstracts of the 43rd ASMS conference on mass spectrometry and allied topics* **1995.**
(13) Lehmann, W. D.; Bohne, A.; von Der Lieth, C. W. *J. Mass Spectrom.* **2000**, *35*, 1335−1341.
(14) Yi, E. C.; Lee, H.; Aebersold, R.; Goodlett, D. R. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2093−2098.
(15) Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R. *Nat. Biotechnol.* **2004**, *22*, 1459−1466.
(16) Keller, A.; Eng, J.; Zhang, N.; Aebesold, R. *Mol. Sys. Biology* [Online] **2005**, DOI: 10.1038/msb4100024.
(17) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.
(18) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646−4658.
(19) Senko, M. W.; Beu, S. C.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229−233.
(20) Anderson NL, Anderson NG. *Mol. Cell. Proteomics* **2002**, *1*, 845. DOI: 10.1074/mcp. R200007-MCP200.
(21) Ghaemmaghami, S.; Huh, W. K.; Bower, K.; Howson, R. W.; Belle, A.; Dephoure, N.; O'Shea, E. K.; Weissman, J. S. *Nature (London)* **2003**, *425*, 737−741.
(22) Ramsey, R. S.; Goeringer, D. E.; McLuckey, S. A. *Anal. Chem.* **1993**, *65*, 3521−3524.
(23) Kast, J.; Gentzel, M.; Wilm, M.; Richardson, K. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 766−776.
(24) Ewing, B.; Green, P. *Genome Res.* **1998**, *8*, 186−194.

PR050436J