Published in final edited form as:

J Proteome Res. 2013 July 5; 12(7): 3207–3214. doi:10.1021/pr4003382.

Use of Theoretical Peptide Distributions in Phosphoproteome Analysis

Mridul Kalita[§], Takhar Kasumov[#], Allan R. Brasier[§], and Rovshan G. Sadygov^{§,*}

[§]Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX 77573

[§]Sealy Center for Molecular Medicine, University of Texas Medical Branch, Galveston, TX 77573

[§]Institute for Translational Sciences, University of Texas Medical Branch, Galveston, TX 77573

[#]Department of Gastroenterology and Hepatology, Cleveland Clinic, 9500 Euclid, Avenue, Cleveland, OH 44195

Abstract

The high mass accuracy and resolution of modern mass spectrometers provides new opportunities to employ theoretical peptide distributions in large-scale proteomic studies. We used theoretical distributions to study noise filtering, mass measurement errors and to examine mass-based differentiation of phosphorylated and non-phosphorylated peptides. Only the monoisotopic mass of the experimental precursor ion was necessary for this analysis. We found that peak deviations can be used to characterize the modification states of peptides in a sample. When applied to large-scale proteomic datasets, the peak deviation distribution can be used to filter chemical/electronic noise for singly charged species. Using peak deviation distributions, it is possible to separate the phosphorylated peptides from the non-phosphorylated peptides, enabling evaluation of the phosphoproteome content of a sample. Because this approach is simple, with light computational requirements, the analysis of theoretical peptide distributions has a significant potential for application to phosphoproteome analyses. For our studies we used publicly available datasets from three large-scale proteomic studies.

Keywords

high mass accuracy informatics; phosphopeptides; theoretical peptides; theoretical phosphopeptide enrichment; tandem mass spectrometry; noise filtering; mass defect

Introduction

The distribution of theoretical peptide masses is structured with well-defined peaks (mass regions densely populated by peptides) and troughs (regions with low peptide densities).¹ For small peptide masses, the troughs contain so-called “forbidden zones” -mass regions that

*To whom correspondence should be addressed. Phone: +1 409 772 3287. Fax: +1 409 772 9679. rovshan.sadygov@utmb.edu.

Note: The authors declare no competing interests.

Supporting Information Available: Supplementary materials include peaks and 95% peak widths (holding 95% of all peptides of a peak) for the phosphorylated (Peaks.Widths.Phospho.txt) and non-phosphorylated (Peaks.Widths.txt) peptides, the number of sequences for each mass bin (up to 3.5 kDa) for the phosphorylated (PhosphorylatedAndNonphosphorylated.Sequences.txt) and non-phosphorylated (NonPhosphorylated.Sequences.txt) peptides, Figures S1-S6 and Table S1. The files of peptide distributions are fixed-width text files. This material is available free of charge via the Internet at <http://pubs.acs.org>. Examples of theoretical peak distributions for short mass intervals, and results of a study of peak width and skewness are available in our earlier publication.¹

do not hold peptide species.² As the peptide mass increases, the forbidden zones shrink in size and finally disappear. However, the peak-centered structure of the peptide masses persists for all mass values. The positions of the peaks evolve on the mass axis, as the distance between the peaks is about 1.00044 Da on average. The high mass accuracy and resolution of modern mass spectrometers creates opportunities for employing the structure of peptide distributions to aid the analyses of mass spectral data from proteomic studies.³⁻⁵ Previous studies were mainly conducted using the “mass defect” (MD), the difference between a monoisotopic peptide mass and its nominal mass, to design chemical tags with unique mass defects,⁵ to separate non-phosphorylated and phosphorylated peptides,⁶ and to filter noise in the mass spectra of proteins obtained via matrix-assisted laser desorption/ionization mass spectrometry.⁷ Interestingly, the latter study found that filtering non-peptide signals increased protein identification scores by up to four orders of magnitude. To generate the mass distributions of theoretical peptides, most studies used peptide sequences from protein sequence databases. By contrast, a phosphopeptide study⁶ analyzed the differences between the mass distributions of all theoretically possible tryptic peptide species (up to a mass of 2 kDa) to identify mass regions where the distributions of non-phosphorylated and phosphorylated peptides (at Ser, Thr and Tyr) were distinct. It was found that for peptide masses up to 1 kDa, more than 98% of all theoretical monophosphorylated peptides can be distinguished from non-phosphorylated peptides directly by their mass. The proportion of separable peptides decreased to 70% at a peptide mass of 1.5 kDa. These observations create possibilities for increasing phosphopeptide sampling by targeting specific mass values, as opposed to the current data-dependent sampling employed in peptide sequence analyses using liquid-chromatography/tandem mass spectrometry (LC-MS/MS). A similar previous study used mass distributions of peptides from an international protein index (IPI) database.⁸ By comparing the mass defects of tryptic peptides derived from human proteome with those of the corresponding phosphopeptides, it was shown that regions of higher phosphorylation probability exist. The conclusions were reported to be relatively insensitive to a mass measurement accuracy of up to 20 ppm. It was suggested that the approach can be used with tandem mass spectrometers in real time to rapidly select and rank order the possible phosphopeptides from a mixture of non-phosphorylated peptides for subsequent phosphorylation site mapping and peptide sequence analysis. Similar observations were found in a study of theoretical glycopeptides.^{9,10} In the present work we examine the use of mass distributions of theoretical peptides to analyze datasets from real experiments. Our study addresses three questions: 1) Does the use of experimental precursor mass information allow estimation of the phosphopeptide content in mass spectral data? 2) Does the use of mass distribution improve the accuracy of precursor mass measurements? and 3) Does the use of mass distribution information allow noise filtering of precursors? Evaluating the phosphoproteome content of a sample is an important issue in phosphoproteomics.¹¹ We will show that a rapid analysis of raw datasets using theoretical peptide distributions allows the estimation of their phosphoproteome content.

Data and Methods

Mass Spectral Data

To perform evaluations of the three potential applications for high mass accuracy information, we used mass spectral data from two publicly available datasets^{12,13} and a dataset generated at the Case Western University.¹⁴ To evaluate the relevance of theoretical peptide distributions in phosphoproteome studies, we used a publicly available phosphoproteome dataset for a mouse brain.¹² From this dataset we used all 10 strong cation exchange (SCX) fractions that were analyzed using collision-induced dissociations (CID). The data were acquired on a LTQ-Orbitrap Velos mass spectrometer. To examine the possible improvement in experimental mass accuracy, we used data from a large-scale sampling of a human proteome.¹³ These are proteome samplings of 11 commonly used

cancer cell lines. Each cell line was fractionated into 6 fractions using strong anion exchange (SAX) and each fraction was analyzed using LTQ-Orbitrap Velos¹⁵ and LTQ-Orbitrap Elite¹⁶ (twice, with a different resolution for MS/MS) mass spectrometers. We report the results for the datasets from all six fractions of the MCF7 cell line that were analyzed with an LTQ-Orbitrap Elite (The results for other cell lines were similar and are not reported here). Note that peptide fragmentation was performed with higher energy collisional dissociation,¹⁷ in contrast to the CID that was used in the murine brain phosphoproteome study. As will be seen below, the fragmentation type is irrelevant to our approach. We use only precursor mass-to-charge ratio and charge state from experimental data. Testing the potential of high mass accuracy for chemical noise filtering requires datasets where precursors with several different charge states are available. It is known that the chemical noise is mostly comprised of solvent molecules with low (mostly +1) charge states.¹⁸ In the acquisition of the MS/MS spectra in the above described datasets, the +1 charged precursors were omitted from the analysis. In the Case Western dataset +1 charged precursors were retained. The latter dataset was obtained from rat heart mitochondria. Note that the complete dataset includes several time points for labeling by D₂O water. We used data only from +1 charged monoisotopic precursors. The mass spectral data were acquired using an LTQ-FTICR mass spectrometer.¹⁹ The datasets for both the cell line and the murine phosphoproteome were downloaded from the Tranche data repository.²⁰ The datasets and their use in our analyses are summarized in Table S1 of the Supplementary Materials section. In all datasets, the mass spectral data were acquired in a data-dependent mode, whereby every survey scan was followed by several tandem mass spectra of precursor ions (Figure S1 of Supplementary Materials section). Abundance-guided heuristics were used to choose the precursor ions for fragmentation. Our program for extracting peak deviations works with the mzML file format.²¹ ThermoFisher Scientific's Raw files were converted into mzML files using Proteowizard's²² msconvert tool. To determine peaks in the distributions of theoretical peptides we used our recently developed algorithm for generating the amino acid compositions and sequences of all theoretically possible peptides with a given enzyme specificity.^{1,23}

In general, and particularly for phosphopeptide content estimation, only experimental precursor mass-to-charge ratios and charge states are used in our approach. To compare our results with those from database searches, we used peptide identification using tandem mass spectra and protein sequence databases.²⁴ Peptide identifications were performed using the Mascot database search engine.²⁵ The database search parameters were: peptide mass accuracy of 20 ppm; variable phosphorylation modifications for Ser, Thr and Tyr; variable oxidation for Met; and fixed carbamidomethylation for Cys. A trypsin enzymatic specificity with one missed cleavage site was used. The identified peptides were filtered with expectation value = 0.05. The database search results were converted into the mzid file format provided in Mascot. All figures and statistical estimates were made using the R environment for statistical computing.²⁶

Modeling Peak Deviations

The mass distributions of peptides are discrete and clustered around peaks that are spaced approximately 1 Da apart. Their widths vary with the mass, and for large masses, increase linearly. For practical peptide mass values up to 3 kDa, we can assume that the observed peptide masses are realized sample points from a single distribution of peak deviations defined below. In our approach every mass value observed in a full mass scan “belongs” to a certain peak of theoretical peptides. The theoretical peak is determined based on the experimental precursor mass – as the peak that is the nearest (in terms of mass difference) to the experimental precursor mass (Supplemental Figure S2). We define the peak deviation

(PD) as the mass difference between the theoretical peak and the monoisotopic mass of the experimental precursor:

$$PD = M(\text{Peak Mass}) - M(\text{Monoisotopic Mass of Experimental Precursor}) \quad (1)$$

The PD can be either positive or negative. Since peaks occur at approximately 1 Da intervals, PD values are limited to the interval [-0.5; 0.5]. Figure S2 of the Supplementary Materials section provides a pseudocode for computing the PD. Peak masses of theoretical peptide distributions, $M(\text{Peak Mass})$, are included in the Supplementary Materials, Peaks.Widths.Phospho.txt for the phosphopeptides and Peaks.Widths.txt for the non-phosphorylated peptides. The PD is a random variable (RV). All experimental observations, R , of a mass difference between the nearest theoretical peak and the monoisotopic mass of experimental precursor can be considered as samplings (realizations) from the PD distribution. For theoretical peptides, the peak deviation, X , has a mean value of zero. If there were no errors in experimental precursor mass measurements, then R would have the same distribution as X . However, experimental precursor masses are determined with non-zero error, S , which itself is a random variable, and is often approximated by a normal distribution, $N(\mu_S, \sigma_S^2)$. μ_S and σ_S are the mean and variance of S , respectively. The experimental measurement of peak deviation, R , can then be modeled as the sum of two independent random variables, one originating from a theoretical peak deviation, X , and the other from the mass measurement error, S :

$$R = X + S \quad (2)$$

According to the linearity of the expectation, the expectation of the sum of RVs is the sum of their expectations. The mean of X is zero, so the mean R is equal to that of S :

$$E[R] = E[S] \quad (3)$$

Since X and S are independent, the variance of R is the sum of the variance of the RVs X and S :

$$\text{var}(R) = \text{var}(X) + \text{var}(S) \quad (4)$$

As a result of experimental mass determinations (via measurements of the mass-to-charge ratio and charge state of an experimental precursor) we have sampled R and can approximate its mean and variance from the sample's mean and variance. The distribution of X is known, as it is that of a theoretical PD. Therefore, from Eqs. (3) and (4) we can estimate the characteristics of the mass error, S - its mean and variance. In principle, these equations provide a path to determining μ_S and σ_S from measured mass values and the known distribution of theoretical peaks *before* doing a database search to identify peptide sequences. For practical applications of these equations the distribution parameters (means and variances) of X and S should be comparable.

Mass Accuracy

There is no single definition of mass accuracy in proteomics.²⁷⁻²⁹ Of particular relevance to our work is the mass accuracy error of a single analyte and the statistical distribution of the errors of all analytes. We determine the mass accuracy error of a single analyte as the difference between the experimental precursor mass and the theoretical mass of the identified peptide sequence (expressed in ppm's). In a typical large-scale proteomic study,

the same analyte may be detected several times over its elution/migration profile and yields a distribution of mass measurement errors. Statistical processing can potentially be used to improve mass measurement errors (accuracy and precision) for a single feature.²⁸ In this work we do not account for the elution time or charge state dependence of mass accuracy errors. Instead, each observed analyte is characterized by the ppm mass error determined from the corresponding peptide sequence assignment. The whole dataset is represented by the statistical distribution of mass errors of individual analytes. Two metrics are inferred from this distribution, sample estimate of the mean of the errors (mass accuracy of all precursor mass measurements) and sample estimate of the variance of the errors (reciprocal of the precision). These two parameters are used in the analyses below (for S).

Results and Discussions

The deviations of monoisotopic masses of experimental precursors from theoretical peaks form a well-defined, centered peak. This is illustrated in Figure 1, which shows the distributions of the peak deviations of all experimental precursors in the first fractions of SAX separations (MCF7 cell line) acquired with three different mass spectrometer settings (mainly relating to the resolution of MS/MS spectra). Every distribution is the result of a single experimental run (single Raw file). We have only used monoisotopic masses of experimental precursors and the positions of peaks from theoretical peptide distributions to generate the figure. As seen from the figure, the distribution of the peak deviations of all experimental precursor masses from one experiment has a single, well-defined mode. The distributions are the realizations of the random variable R in Eq. (2) in the specific experiments (three replicates of the 1st fraction of the MCF7 cell line). The observed distribution of PD is quite different from that of MD for the same experiments. The distributions of MD are rather broad with multiple modes (Supplementary Figures S3 and S4). Our observations show that the peak deviation serves as a convenient measure for experimental precursor masses.

The modes of peak deviation distributions in Figure 1 are non-zero. This was expected as the theoretical peptides were generated with equal proportions of all amino acids. In the human proteome (like the proteomes of other species) certain amino acids are more abundant (for example, leucine in the human proteome). This difference in amino acid compositions creates a relative shift in the distributions of theoretical and natural protein peptides. The important point from the figure is that the PDs of experimental precursor masses form a single, well-defined peak. Sampling peptides in a data-dependent acquisition mode is now transformed into a sampling from this single, well-defined peak. The rest of the discussion in this section addresses ways of employing the distribution in processing datasets from large-scale proteomic experiments to obtain information about the sample content before performing database searches.

When we created distributions for different separation fractions of the murine brain phosphopeptide dataset,¹² it was clear that the distributions reflected changes in the phosphopeptide content from fraction to fraction, as seen in Figure 2. The figure shows the results from the 1st (blue), 5th (red) and 8th (black) SCX fractions. The PDs were computed from mass spectral data (monoisotopic masses of experimental precursors from three Raw/mzML files) and theoretical peaks only – no information from any database search was used. The shapes and modes of PD distributions are clearly dependent on the fraction composition. Previous studies^{6,30} have shown that mass defects of phosphorylated peptides are larger than those of the non-phosphorylated peptides. We associated the blue peak (1st fraction) with phosphorylated peptides, and the black peak (8th fraction) with non-phosphorylated peptides. The relative positions of species-specific peaks depend on the definition of the peak deviation. We have defined the PD as the difference between the

nearest theoretical peak position and the monoisotopic mass of experimental precursor. Based on this definition, the PDs of phosphorylated peptides are positive, and those of non-phosphorylated peptides are negative. While the PDs of the 1st and 8th fractions seemingly show unimodal distributions, there are two visible modes in the PDs of the 5th fraction (shown in red). This behavior indicates that early fractions were more enriched in phosphopeptides than the later ones. The 5th fraction exhibits a bimodal distribution with two identifiable peaks corresponding to non-phosphorylated and phosphorylated peptides. Similar results were observed for other SCX fractions (Supplemental Figures S5 and S6).

Peptide identifications from database searches can be used for comparisons with our observations. We parsed the database search results from the 5th fraction for peptides containing at least one phosphorylation site and non-phosphorylated peptides. For each of these two groups, we generated the density of peak deviations. In Figure 3, we plotted the peak deviations of all (solid lines), phosphorylated (dotted line) and non-phosphorylated peptides (dashed line). As seen from the figure, the two major modes in the density profile of the unfiltered data overlap well with the modes of phosphorylated and non-phosphorylated peptides. We note here that while the relative peak heights on the same density curve (solid line) can be used to estimate the relative ratios of the phosphorylated and non-phosphorylated peptides, density profiles from different distributions are not relevant for ratio estimations.

From Figure 2 it is seen that relative peak height information can be obtained from the distributions of PDs for fractions with two distinguishable peaks. We have used an unsupervised classification method, k-means clustering,²⁶ to evaluate the number of non-phosphorylated (initial value of PD peak centered at -0.1) and phosphorylated (initial value of PD peak centered at 0.1) peptides. For six of the ten fractions the k-means clustering could identify two peaks associated with the phosphorylated and non-phosphorylated peptides (based on the peak center location). Figure 4 shows the relative proportions of the phosphorylated and non-phosphorylated peptides as a function of fractionation steps. The PD distributions for the first three fractions and the tenth fraction are unimodal, with undetectable peaks (by the k-means algorithm) – for non-phosphorylated peptides (fractions 1 through 3) or for phosphorylated peptides (10th fraction), as illustrated in Figures S5 and S6. The phosphopeptide content dynamics in Figure 4 (computed solely using the monoisotopic masses of experimental precursors and theoretical peptide distributions) correlates well with the results from database searching (Figure 3B in Jedrychowski et al.)¹² In agreement with that figure, our Figure 4 shows that the smallest proportion of phosphopeptides is predicted to be in the 6th fraction, and there is a gradually increasing phosphopeptide content in the 7th and 8th fractions followed by a decline in the 9th fraction. Among the fractions shown in Figure 4, the 4th fraction is most enriched in phosphopeptides (again, in agreement with Jedrychowski et al.)¹² Thus the simple calculations of peak deviations (using only the peaks in theoretical peptides distribution and monoisotopic masses of experimental precursors, and requiring very little computational resources) serves as an alternative to database searching for evaluating the phosphoproteome content of a sample. The use of peak deviations may also make it possible to evaluate the efficiency of phosphopeptide analysis, by comparing the ratio of the number of identified phosphopeptides to the number of identified non-phosphorylated peptides with the corresponding ratio obtained from the distribution of peak deviations. If the efficiency of identifying phosphopeptides and non-phosphorylated peptides were similar, then the ratio of the number of identified phosphopeptides to the number of identified non-phosphorylated peptides should correlate with the ratios of the corresponding peaks observed in the distribution of peak deviations (Figure 4). We note that all figures related to the phosphoproteome study (Figures 2-4, S4 and S5) included all possible phosphorylation

states (mono-, di-, ... phosphorylated peptides); there was no filtering with respect to the number of phosphorylated sites.

Evaluating the phosphoproteome content of a peptide mixture is relevant for such purposes as optimizing enrichment, fractionation and chromatographic separation – procedures that are essential in phosphoproteomics.^{31,32} Different combinations of sample preparation procedures are optimal for varying amounts of a starting peptide mixture. Peptide identification via database searching is normally employed to evaluate the phosphoproteome content of a peptide mixture.^{12,32} Our approach, in which we use only the monoisotopic masses of experimental precursors, offers an alternative strategy wherein evaluation of the phosphoproteome content does not depend on the efficiency of phosphopeptide fragmentation and database search algorithms used for identifying peptide sequences.

Next we compared the distribution of peak deviations with that of mass measurement errors for monoisotopic masses of experimental precursors. It has been suggested that the theoretical peaks could be used to correct the mass measurement errors for product ions.³³ In that study, theoretical corrections to the fragment ion mass-to-charge ratio values somewhat improved the peptide identifications from database searches. This is a potentially attractive application of theoretical peptide mass distributions. Figure 5 shows a comparison of the experimental precursor mass error distribution and the corresponding peak deviations from theoretical peptide distributions. For this figure we used the mass spectral data¹³ from the 1st fraction of the MCF7 cell line acquired on an Orbitrap Velos mass spectrometer. Figure 5 is illustrative of all other fractions and replicates (data not shown). The distribution of peak deviations (black line) requires no information from peptide identifications using database searches. Experimental mass measurement errors (blue line) were determined using the measured precursor masses and the calculated masses of peptide sequences identified via tandem mass spectra and database searching. To better illustrate the distributions, we used two different sets of x- and y-axes (with substantially differing scales). As seen from the figure, the spread (variance) of the distributions is substantially different – the standard deviations of mass measurement errors and peak deviations are 5 ppm and 45 ppm, respectively. The large difference (almost an order of magnitude) between the standard deviations seems to indicate that improving the accuracy of mass measurements, for high resolution and mass accuracy measurements, by using theoretical peptide distributions is not feasible. 45 ppm translates to 0.45 Da at 1 kDa. In cases where ions are measured on ion traps, the standard deviations of mass errors and theoretical peptide distributions are similar – 0.5 Da and 0.45 Da, respectively; therefore, some correction could potentially be obtained using theoretical peptide distributions, as was reported by Egertson et al.³³

Figures 6A and B plot the distribution of Mascot ion scores vs. peak deviations for the murine phosphoproteome and the MCF7 cells, respectively. As seen from the figures, for data points on the tails of the peak deviation distributions, the Mascot scores are zero. However, these data points constitute only a small portion of the overall data. Thus, the combined peak deviations < -0.2 or > 0.4 represent 260 sample points (1% of all data points) in the murine phosphoproteome dataset (Figure 6 A). Of these only six had ion scores above 20. For the MCF7 cell line, the spread of the PD distribution was larger; only those < -0.3 Da or > 0.2 Da held data points that could be filtered, but they constituted only a small portion of the data (less than 1%). Note that these datasets were obtained in an acquisition mode, where +1 charged species were excluded from fragmentation. It is known that most of the chemical noise represents small organic molecules in the +1 charge state.¹⁸ We therefore examined another dataset where +1 charged species were included in the fragmentation and thus available for our peak deviations analysis. This was a dataset acquired for studies of protein turnover in rat heart mitochondria from different cellular compartments (the Case Western dataset). The results are shown in Figure 7. There were more than 5200 tandem

mass spectra of +1 charged experimental precursors in this dataset. The plot of Mascot ion scores vs. peak deviations shows clearly identifiable regions of peak deviations where the ion scores are very small (less than 5). A conservative filtering (removing peak deviation values < -0.15 or > 0.1) filtered 35% of all precursors with no effect on the identification results. We believe that most of these data points originate from chemical noise whose peak deviation distribution is substantially different from that of the peptide species. The observation suggests that the peak deviations can be used to filter out chemical noise (in this case, +1 charged).

In summary, we studied the properties of a new quantity, peak deviation, defined as the mass difference between the peaks of theoretical peptide distributions and the monoisotopic masses of experimental precursors. The peak deviation is different from the more familiar mass defect calculated using monoisotopic and nominal masses of experimental precursors. Our analyses of several large-scale datasets showed that peptide sampling can be modeled as a sampling from a distribution of peak deviations. We then used the properties of the peak deviation distribution to distinguish between phosphorylated and non-phosphorylated peptides, to examine possible improvements to the mass measurement errors and to perform noise filtering of experimental precursors. To the best of our knowledge this is the first direct application of theoretical peptide distributions to an experimental dataset from phosphoproteomics. In future work we will model the peak deviation distributions of phosphorylated and non-phosphorylated peptides with the aim of improving confidence in phosphopeptide identifications.

Conclusions

In this work we have examined three potential applications of peak deviations to high throughput proteomic experiments. We have shown that peak deviations are useful to estimate the phosphoproteome content of a complex peptide mixture. This stems from the fact that phosphorylated and non-phosphorylated peptides have different peak deviation distributions. Samples containing a mixture of these peptides exhibit a bimodal peak deviation distribution. The relative amounts of the modification states can be inferred from deconvolution of the mixture distribution. It is important to note that our analysis does not require a database search to estimate the phosphoproteome content; all this estimation requires are the peaks of theoretical peptides and the monoisotopic masses of experimental precursors.

We found that the standard deviation of the experimental mass measurement errors is substantially smaller than that of the peak deviation distribution. Therefore, improving the experimental mass accuracy for the modern mass spectrometers (~ 5 ppm) is not straightforward, if possible at all, using theoretical peptide distributions. Our analyses using peak deviations for noise filtering showed that the peak deviations are helpful for filtering out +1 charged noise.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported, in part, by UL1TR000071, the UTMB Clinical and Translational Award (ARB); HHSN272200800048C, the NIAID Clinical Proteomics Center (ARB); and NIH-NLBIHHSN268201000037C, the NHLBI Proteomics Center for Airway Inflammation (Alex Kurosky, UTMB). We thank Dr. William C. Stanley from the University of Sydney, Sydney, Australia, for providing cardiac mitochondrial samples of D₂O-metabolic labeling experiments in rats. We thank Dr. David Konkel for critically editing the manuscript.

References

1. Nefedov AV, Mitra I, Brasier AR, Sadygov RG. Examining troughs in the mass distribution of all theoretically possible tryptic peptides. *J Proteome Res.* 2011; 10:4150–57. [PubMed: 21780838]
2. Mann, M. Useful Tables of Possible and Probable Peptide Masses; Annual Conference on Mass Spectrometry and Allied Topics; 5-1-1995; Atlanta, GA: American Society of Mass Spectrometry; Nov 5. 1995 Ref Type: Conference Proceeding
3. Frahm JL, Howard BE, Heber S, Muddiman DC. Accessible proteomics space and its implications for peak capacity for zero-, one- and two-dimensional separations coupled with FT-ICR and TOF mass spectrometry. *J Mass Spectrom.* 2006; 41:281–88. [PubMed: 16538648]
4. Hernandez H, Niehauser S, Boltz SA, Gawandi V, Phillips RS, Amster IJ. Mass defect labeling of cysteine for improving peptide assignment in shotgun proteomic analyses. *Anal Chem.* 2006; 78:3417–23. [PubMed: 16689545]
5. Yao X, Diego P, Ramos AA, Shi Y. Averagine-scaling analysis and fragment ion mass defect labeling in peptide mass spectrometry. *Anal Chem.* 2008; 80:7383–91. [PubMed: 18778085]
6. Spengler B, Hester A. Mass-based classification (MBC) of peptides: highly accurate precursor ion mass values can be used to directly recognize peptide phosphorylation. *J Am Soc Mass Spectrom.* 2008; 19:1808–12. [PubMed: 18804385]
7. Dodds ED, An HJ, Hagerman PJ, Lebrilla CB. Enhanced peptide mass fingerprinting through high mass accuracy: Exclusion of non-peptide signals based on residual mass. *J Proteome Res.* 2006; 5:1195–203. [PubMed: 16674109]
8. Bruce C, Shifman MA, Miller P, Gulcicek EE. Probabilistic enrichment of phosphopeptides by their mass defect. *Anal Chem.* 2006; 78:4374–82. [PubMed: 16808444]
9. Lehmann WD, Bohne A, von Der Lieth CW. The information encrypted in accurate peptide masses-improved protein identification and assistance in glycopeptide identification and characterization. *J Mass Spectrom.* 2000; 35:1335–41. [PubMed: 11114093]
10. Froehlich JW, Dodds ED, Wilhelm M, Serang O, Steen JA, Lee RS. A classifier based on accurate mass measurements to aid large scale, unbiased glycoproteomics. *Mol Cell Proteomics.* 2013; 12:1017–25. [PubMed: 23438733]
11. Palumbo AM, Smith SA, Kalcic CL, Dantus M, Stemmer PM, Reid GE. Tandem mass spectrometry strategies for phosphoproteome analysis. *Mass Spectrom Rev.* 2011
12. Jedrychowski MP, Huttlin EL, Haas W, Sowa ME, Rad R, Gygi SP. Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics. *Mol Cell Proteomics.* 2011; 10:M111. [PubMed: 21917720]
13. Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics.* 2012; 11:M111.
14. Li L, Willard B, Rachdaoui N, Kirwan JP, Sadygov RG, Stanley WC, Previs S, McCullough AJ, Kasumov T. Plasma proteome dynamics: analysis of lipoproteins and acute phase response proteins with 2H₂O-metabolic labeling. *Mol Cell Proteomics.* 2012
15. Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, Lange O, Remes P, Taylor D, Splendore M, Wouters ER, Senko M, Makarov A, Mann M, Horning S. A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics.* 2009; 8:2759–69. [PubMed: 19828875]
16. Michalski A, Damoc E, Lange O, Denisov E, Nolting D, Muller M, Viner R, Schwartz J, Remes P, Belford M, Dunyach JJ, Cox J, Horning S, Mann M, Makarov A. Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol Cell Proteomics.* 2012; 11:O111. [PubMed: 22159718]
17. Olsen JV, Macek B, Lange O, Makarov A, Horning S, Mann M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods.* 2007; 4:709–12. [PubMed: 17721543]
18. Michalski A, Cox J, Mann M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC-MS/MS. *J Proteome Res.* 2011; 10:1785–93. [PubMed: 21309581]

19. Syka JE, Marto JA, Bai DL, Horning S, Senko MW, Schwartz JC, Ueberheide B, Garcia B, Busby S, Muratore T, Shabanowitz J, Hunt DF. Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J Proteome Res.* 2004; 3:621–26. [PubMed: 15253445]
20. Falkner JA, Andrews PC. P6-T Tranche: Secure Decentralized Data Storage for the Proteomics. *Community Journal of Biomolecular Techniques.* 2007; 18:1.
21. Gilski MJ, Sadygov RG. Comparison of Programmatic Approaches for Efficient Accessing to mzML Files. *J Data Mining Genomics Proteomics.* 2011; 2
22. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics.* 2008; 24:2534–36. [PubMed: 18606607]
23. Nefedov AV, Sadygov RG. A parallel method for enumerating amino acid compositions and masses of all theoretical peptides. *BMC Bioinformatics.* 2011; 12:432. [PubMed: 22059886]
24. Eng JK, McCormack AL, Yates JR. III An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J Am Soc Mass Spectrom.* 1994; 5:976–89.
25. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis.* 1999; 20:3551–67. [PubMed: 10612281]
26. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2010. Ref Type: Computer Program
27. Zubarev R, Mann M. On the proper use of mass accuracy in proteomics. *Mol Cell Proteomics.* 2007; 6:377–81. [PubMed: 17164402]
28. Liu T, Belov ME, Jaitly N, Qian WJ, Smith RD. Accurate mass measurements in proteomics. *Chem Rev.* 2007; 107:3621–53. [PubMed: 17649984]
29. Mann M, Kelleher NL. Precision proteomics: the case for high resolution and high mass accuracy. *Proc Natl Acad Sci U S A.* 2008; 105:18132–38. [PubMed: 18818311]
30. Mitra I, Nefedov AV, Brasier AR, Sadygov RG. Improved mass defect model for theoretical tryptic peptides. *Anal Chem.* 2012; 84:3026–32. [PubMed: 22401145]
31. Nilsson CL. Advances in quantitative phosphoproteomics. *Anal Chem.* 2012; 84:735–46. [PubMed: 22043985]
32. Zhou H, Di PS, Preisinger C, Peng M, Polat AN, Heck AJ, Mohammed S. Toward a comprehensive characterization of a human cancer cell phosphoproteome. *J Proteome Res.* 2013; 12:260–71. [PubMed: 23186163]
33. Egertson JD, Eng JK, Bereman MS, Hsieh EJ, Merrihew GE, MacCoss MJ. De Novo Correction of Mass Measurement Error in Low Resolution Tandem MS Spectra for Shotgun Proteomics. *J Am Soc Mass Spectrom.* 2012; 23:2075–82. [PubMed: 23007965]

Abbreviations

Da	Dalton
CID	collision-induced dissociation
Eq	equation
FTICR-MS	Fourier transform ion cyclotron resonance mass spectrometry
IPI	international protein index
LC	liquid chromatography
LTQ	Linear trap quadrupole
MD	Mass defect
MS/MS	tandem mass spectrometry
MS	mass spectrometry

PD	peak deviation
ppm	parts per million
SAX	strong anion exchange
SCX	strong cation exchange

Figure 1. Distribution of peak deviations from the first fractions of triplicate repeats of whole-cell lysates of the MCF7 cell line.¹³ The peak deviation is computed as the difference between the nearest peak (of the distributions of all theoretically possible tryptic peptides) and the monoisotopic mass of the experimental precursor.

Figure 2.

The distributions of peak deviations for three SCX fractions of phosphopeptides.¹² The black curve denotes the distribution for the 8th fraction, the red curve for the 5th fraction, and the blue curve for the 1st fraction. It is important to note that the figure was generated using raw data – no database search or information about the sequence identity of a peptide was used. The distributions clearly demonstrate the presence of two peaks – one corresponding to the phosphorylated peptides (blue) and one to non-phosphorylated peptides (black). This phenomenon is due to the different mass defects of phosphorylated and non-phosphorylated peptides.⁸ The distributions for other fractions are shown in the Supplementary Materials section (Figures S5 and S6).

Figure 3. Densities of peak deviations of all peptides (solid line), of all phosphorylated peptides (dotted line) and all non-phosphorylated peptides (dashed line) of the 5th fraction.¹² Note that the solid line was generated from the raw data without any information about peptide identities. The broken and dotted lines were generated using Mascot database search results. The densities for phosphorylated peptides were obtained from database identification results containing at least one phosphorylated site.

Figure 4.

Estimated proportions of phosphopeptide content for fractions 4 through 9 from murine phosphoproteomic data.¹² The phosphopeptide content was estimated as a ratio of the sizes of the two clusters obtained by k-means clustering of the peak deviations for these fractions. The phosphoproteome content changes (by fraction number) are in agreement with the results from database searching.¹²

Figure 5. Densities of peak deviation (black) and mass accuracy (blue) for the 1st SAX fraction of MCF cells. The standard deviation of the experimental mass measurement error was 5.2 ppm, while that of the peak deviations was 45 ppm.

Figure 6 A.

Figure 6 B.

Figure 6.

A. Joint distributions of peak deviations and Mascot ion scores for the 5th fraction of the murine brain phosphoproteome (**A**) and the 1st fraction of the MCF7 cell line (**B**). For the phosphoproteome data there is a substantial number of data points on the tails of the peak deviation distribution. These sampling points could be omitted, as they did not produce any reliable Mascot ion scores. The combined peak deviation intervals of less than -0.2 or larger than 0.4 peak deviations hold 260 sample points (% 1 of the total). Of these only six have ion score above 20. For the MCF7 cell line dataset, B, peak deviations less than -0.25 produced an average ion score of 5.

Figure 7.

The joint distribution of Mascot ion scores and peak deviations for +1 charged peptides from a rat heart mitochondrial dataset. There were more than 5200 spectra with +1 charge precursors in this dataset. The union of two peak deviation intervals, those less than -0.15 and those more than 0.1, held 35% of all precursors. There was no identification with a Mascot ion score higher than 5 among these precursors.