Published in final edited form as:
Anal Chem. 2006 July 1; 78(13): 4374–4382. doi:10.1021/ac060046w.

Probabilistic Enrichment of Phosphopeptides by their Mass Defect

Can Bruce[1,*], **Mark A. Shifman**[1], **Perry Miller**[1], and **Erol E. Gulcicek**[2,3]

1*Center for Medical Informatics, Yale University, New Haven, CT 06511*

2*Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511*

3*W.M. Keck Foundation Biotechnology Resource Laboratory, Yale University, New Haven, CT 06511*

Abstract

The mass defect, the difference between the nominal and actual monoisotopic mass of a phosphor in a phosphate group is greater than for most other atoms present in proteins. When the mass defects of tryptic peptides derived from the human proteome are plotted against their masses, phosphopeptides tend to fall off the regression line. By calculating the masses of all potential tryptic peptides from the human proteome, we show that regions of higher phosphorylation probability exist on such a plot. We developed a transformation function to estimate the mass defect of a peptide from its monoisotopic mass and empirically defined a simple formula for a user-selectable discriminant line that categorizes a peptide mass according to its probability of being phosphorylated. Our method performs similarly well on phosphopeptides derived from a database of experimentally validated phosphoproteins. The method is relatively insensitive to mass measurement error of up to 20 ppm. The approach can be used with a tandem mass spectrometer in real time to rapidly select and rank order the possible phosphopeptides from a mixture of unmodified peptides for subsequent phosphorylation site mapping and peptide sequence analysis.

Introduction

Protein phosphorylation plays an important role in intracellular signal transduction [1-4] and is involved in regulating cell cycle progression, differentiation, transformation, development, peptide hormone response, and adaptation. This important post-translational modification often occurs at very low stoichiometry. As a result, it is important that studies directed at identifying phosphoproteins and characterizing phosphorylation sites by mass spectrometry incorporate a phosphoprotein [5, 6] or phosphopeptide enrichment step [7-15] and implement relevant tandem mass spectrometry detection techniques [16-22] that are directed specifically in selecting and identifying phosphopeptides.

Due to its low stoichiometry and lower ionization efficiencies, when using a tandem mass spectrometer instrument directed for phosphopeptide detection, it is difficult to determine if a particular peptide is phosphorylated in the “survey mode” of operation. As a result, valuable time is spent using the mass spectrometer for sequence determination of unphosphorylated peptides that are generally the most abundant peptides. This time commitment may be in excess of the time window in which a transient species may be available for analysis, particularly if

* To whom correspondence should be addressed: Can.Bruce@yale.edu.

Supplementary Material Available: A listing of tables detailing the composition of the tryptic peptides used in this study, and characterization of peptides below and above the $L_{0.9}$ line is available as Supporting Information. Current ordering information is found on any masthead page.

the MS/MS is coupled to an online chemical separation technique such as liquid chromatography or capillary electrophoresis. In this study, we use the concept of fractional mass (the value of m/z to the right of the decimal point) as derived from a mass defect to be able to detect phosphopeptides apart from their more abundant unmodified counterparts in a mass spectrometer. The calculations and algorithms described here use theoretical negative mass shift behavior of phosphopeptides and assigns probabilities to these peptides that can be distinctively selected for MS/MS processing before the more abundant unmodified peptides. The idea can be implemented in real time to first rank order probabilities of peptides for being a phosphopeptide and preferentially decide the order of acquisition for MS/MS from a list of possible peptides. This use of the mass defect probabilities would also increase the confidence and the probability of identification of many phosphopeptides, especially if used in conjunction with the phosphopeptide enrichment methods.

The mass defect is defined as the difference between the actual mass of a specific nucleus and its integer, or nominal, mass (i.e., the sum of protons and neutrons). It represents the mass equivalent (in terms of Einstein's concept of mass-energy interchange) of the different binding energies required for nuclear stabilization of the different elements. The mass defect of ^{12}C is defined as zero atomic mass units (amu) by convention. Without any modifications, the most common naturally occurring 20 amino acids in proteins contain the elements C, H, N, O and S. On such a scale, the mass defects of other elements commonly found in amino acids differ negligibly from that of carbon: ^{14}N is 0.0031 amu, ^{16}O is −0.0051 amu, and ^{1}H is 0.0078 amu. The mass defect for ^{31}S, a much less frequently encountered element in protein, is rather larger, being −0.0279. Similarly, ^{31}P has a negative mass defect (−0.0262) that is significantly larger than most other commonly found atoms in biomolecules. When it is present on proteins in the form of a phosphate group (net addition of HPO_3), the total mass deficit shifts by −0.0337. In this study, we hypothesized that phosphate would act as a natural mass defect tag and might be useful in identification and characterization of phosphorylated peptides.

The use of mass defect principles in mass spectrometry dates back to its early applications by Edward Kendrick[23] in the simplification of analysis of organic compounds by high resolution mass spectrometers. The utility of this application takes advantage of the repeating mass of (CH2) in hydrocarbons, and it is still used very effectively today with very high resolving power mass spectrometers [24-26]. The composition of structurally related complex hydrocarbons from petroleum products or other natural sources can be “classified” according to their mass defects by their two dimensional Kendrick Mass Defect plots [24]. While based on the same mass defect principle, additional concept of graphically depicting elemental composition ratios in two and three dimensional van Krevelen diagrams for elucidating compositional differences of complex organic mixtures between different fossil fuels has been described in detail by Wu et.al. [27]

There has been more recent applications of the mass defect principles, in particular, in the identification and quantitation of proteins [28-35], in the characterization of lipids and phospholipids from intact bacterial cells [36], as well as in its use in identification of metabolite ions [37]. In identification of proteins by their accurate mass, Gay et al.[28] has shown that the clustered distribution behavior of peptides up to mass 3000 in one Dalton intervals can be used to improve Peptide Mass Fingerprinting identification of peptides. Similarly Conrads et al. [29] has demonstrated that if the molecular mass of a peptide can be determined with high enough mass measurement accuracy, its unique mass can be used in the identification of proteins as Accurate Mass Tag (AMT). They also suggest that this unique feature can be used to determine the identification of post translationally modified peptides especially if their modification poses a large enough of a mass defect compared to other abundant elements in peptides. In similar approaches, Amster's group has shown that identification of proteins by accurate mass measurement of their tryptic digested peptides in a MALDI FTMS instrument can be improved

by incorporating a mass defect label containing two bromines into the cysteine residues of proteins [30, 31] The labeled peptides containing cysteines will be shifted by approximately 0.3 Da lower than the average mass defect of the underivatized peptide (around 1800 Da) placing them in a region of the mass spectrum that is not occupied by unmodified peptides. The shift caused by the mass defect will allow the recognition of the labeled peptide making its identification with much higher specificity. In top down identification of proteins, recently Hall *et. al*. have described a technique termed inverted mass ladder sequencing (IMLS) for quickly obtaining an N- or C-terminal protein sequence tags from intact proteins by in-source fragmentation in an electrospray ionization time-of-flight (ESI-TOF) mass spectrometer [32, 33]. IMLS involves labeling the N- or C-terminus of a protein with a large mass defect element such as Bromine or Iodine that allows resultant assembly of peptides to be shifted by the amount of the mass defect and distinguished away from all the untagged peptide species deemed "chemical noise". The ability of these elements as well as the rare earth elements that exhibit large mass defects enables their associated mass tags to be used as affinity enrichment "handles" that can also be used as quantitation tools for specific proteins or classes of proteins of interest from complex mixtures [34, 35]. While the use of elemental compositions which exhibit mass defect has been previously used as above by numerous investigators, the potential for probabilistic phosphopeptide determination for consequent MS/MS among other interfering peptides in the survey mode have not been explored.

To assess the utility of using mass defect in estimating the probability that a peptide might contain a phosphate group, we performed a computational study on currently known elements of the human proteome and on proteins with well characterized phosphorylation sites. Our findings suggest that based purely on the monoisotopic mass value, this approach can effectively identify a subset of peptides as being likely phosphorylated.

Experimental

Dataset creation

The IPI Human protein database [38], version 3.07, containing 50,207 proteins, was used to generate tryptic peptide sequences. Each peptide was allowed to have 0, 1 or 2 missed tryptic cleavage sites (peptides containing unknown amino acids, coded as 'X' in the database, were excluded from the analysis). Cysteines were assumed to be carbamidomethylated. Peptides were allowed to be variably phosphorylated at 1, 2 or 3 potential phosphorylation sites on S, T and Y residues and to be variably oxidized at methionine residues. Monoisotopic masses were calculated for each phosphorylation and methionine-oxidation variant of a peptide by the addition of the appropriate number of HPO_3 and O masses, respectively. Because the resulting peptide set contained multiple occurrences of the same sequences originating from different protein digests, subsequent analyses were performed on a list in which each sequence (and its phosphorylation and methionine-oxidized mass variants) is unique. When all of the above conditions were applied, 16,122,844 unique tryptic peptides were used in the calculations.

In order to generate a set of tryptic peptides with validated phosphorylation sites, the phosphoELM database, v.3.0, was used [39]. The subset of human proteins in this database was analyzed as above to generate a list of validated, phosphorylated, tryptic peptides. The distribution of the various peptide species from the IPI Human, IPI mouse and phosphoELM databases, according to their missed cleavage status and number of potential phosphorylation sites are shown in Table S1.

Estimation of the mass defect of a peptide from its fractional mass

The mass defects of all peptides derived from the IPI database, with 0 and 3 phosphate groups, were separately fitted to linear regression lines. Two nearly parallel lines resulted and the

equation for a new line between the two lines was calculated by averaging their a and b linear regression parameter values, to yield a_m and b_m. Two lines with parameter of a_m and b_m −0.5, and of a_m and b_m +0.5 were taken to be the lower and upper boundaries of a diagonal strip containing fractional mass values.

As mass defect values increase beyond 1, their (experimental) fractional mass "wraps around" back to 0. The mass defect of a peptide can be estimated from its fractional mass, based on the distribution of mass defect values as a function of mass. This can be done by determining the equation of a zigzag-shaped line of the form y=ax + b-int(ax + b) (where the "int" function converts a number to the largest integer less than that number) that separates the upper edge of the fractional mass data cloud from its lower edge, as it wraps around between 0 and 1. We assumed the a and b parameters of this function to correspond to a_m and b_m −0.5 mentioned in the previous paragraph and assigned them the values of a = 0.00050 and b= −0.59.

The sequence and mass of tryptic peptides was determined with a program written in Perl. Regression analysis was preformed with the plotting program gnuplot, other statistical tests with R.

Results and Discussion

We used two different types of data sets to analyze the distribution of singly charged (MH^+) monoisotopic m/z values for phosphopeptides. One is derived from the IPI Human protein database, a comprehensive and non-redundant proteomics database based on genomic sequence. The other is derived from proteins found in the phospho.ELM database, which lists proteins with experimentally validated phosphorylation sites [39]. The data from the IPI database contains every potential phosphorylated sequence from the currently known human proteome, but we do not know which of these are actually ever phosphorylated. The dataset derived from the phospho.ELM database contains actual phosphopeptides sequences, but limited in size and its bias towards proteins in well-studied systems.

Because we allowed for peptides to have up to 2 missed tryptic cleavage sites, some peptides were large enough to have multiple phosphorylation sites. In the dataset derived from the validated phosphoproteins database, 72.4, 18.0, 4.8 and 4.8% of the phosphopeptides had 0, 1, 2, 3 and >3 phosphate groups, respectively. Because the vast majority of validated phosphopeptides had 3 or fewer phosphates, we chose to limit our analysis of peptides from the IPI database to those that had up to 3 phosphates.

The relation of mass defect to mass is illustrated as a histogram for peptides having a mass within the range of 1200−1201 Da (Figure 1). Here, for unphosphorylated peptides, the fractional mass ranges between 0.4 and 0.8. For phosphorylated peptides, the ranges of fractional masses are shifted to lower values according to the number of phosphate groups they carry. Each peptide class has a mean fractional mass that is ~0.09 Da. lower than the class containing one fewer phosphate. Peptides with a mass between 1200.0 and 1200.4 are nearly certain to be phosphorylated when compared to unmodified peptides. The probability that a peptide in this population is phosphorylated can be estimated by dividing the number of phosphorylated peptides to all peptides having a mass within a given mass range. To estimate the probability of phosphorylation we used the formula (N1+N2+N3)/(N0+N1+N2+N3) where N0, N1, N2 and N3 refer to the number of peptide sequences having 0, 1, 2, and 3 phosphates, respectively.

The distribution of peptides as a function of mass defect and monoisotopic, MH^+, mass can be visualized with a color map shown in Figure 2. In this figure, each pixel represents a bin that is 10 Da monoisotopic mass wide and 0.01 Da mass defect high, and the number of peptides within that bin is color-coded as indicated. Panels A, B, C, and D show the distribution of

tryptic peptides phosphorylated at 0, 1, 2 and 3 sites, respectively. The green contour in Panel A shows where the number of unphosphorylated peptides per bin is 10 or more (96% of such peptides are within this area). This same contour area is displayed in the four panels for ease of comparison among them. Thus, the histograms in Figure 1 represents a cross-section of the data in the panels A, B, C, D and E of Figure 2, where monoisotopic mass = 1200 Da. It can be seen that there is a linear relationship between the fractional mass and mass for peptides. The slope of the linear regression lines (in the form y = ax+b) for each panel are about equal, while the intercept values shift down by about 0.08 Da with the addition of each additional phosphate group (Table S2). We also performed this analysis on peptides derived from proteins of the Mouse IPI database and found that the mouse IPI database proteins had regression lines with essentially identical parameters as the parameters for data derived from the human IPI database (Table S2). For the validated phosphopeptides derived from the phospo.ELM database, the linear regression lines also shifted down for each additional phosphate group by a similar amount, although the regression line parameters were slightly different from the lines of the other two data sets examined (Table S2).

In all cases, singly phosphorylated peptides are more abundant in numbers than doubly and triply phosphorylated peptides and have a greater overlap with unmodified peptides. But the shifts in mass defect still presents a significant statistical opportunity to make better decisions while performing shotgun analysis which otherwise would be absent.

The downward shift in the distribution of phosphorylated peptides in these density maps results in the lower parts of the probability map having a high value for phosphorylation probability. It is clear from the mass defect plot (Fig 2E) that there is a region appearing as a yellow diagonal strip where the probability of a peptide being phosphorylated is greater than 90%.

In the mass range for the data shown in Figure 2, because mass defect values are less than 1, they equal fractional mass, the part of the mass value that is to the right of the decimal point. However, as mass defect values increase beyond 1 with increasing molecular mass, fractional mass has to cycle back to 0. In order to make useful determination of phosphorylation probabilities of peptides with larger molecular mass beyond this 0-to-1 cycling point, observed fractional mass has to be correlated to the theoretical mass defect. The point of this exercise would be to define a simple function that expresses the boundary of a region with high phosphorylation probability. This process can be seen in Fig. 3, where number of peptides is first mapped as a function of their fractional mass and monoisotopic mass (Fig. 3A, blue and purple lines). The blue (no phosphate) and purple (3 phosphates) contour lines defining the region with 10 or more peptides per pixel (1 pixel represents a bin of 10 Da mass × 0.01 fractional mass or mass defect) for two types of peptides zigzags between 0 and 1 as peptide mass increases (the number of peptides outside each contour line represent 4% of the total). Next, two lines with the formulas y=ax+b-int(ax+b) and y= 1+ax+b-int(ax+b) were fitted such that when they are drawn on a graph plotting fractional mass against mass, the first line grazes the lower edge of the data cloud and the second one the upper edge (dashed lines in Fig 3A). As described in the Methods, we assumed that for the Mass Defect vs. Monoisotopic Mass data, that a line of formula y=0.5+ax+b approximates the regression line going through data points corresponding to all peptides. As can be seen in Figure 3A, most of the data is bound between two continuous lines y=ax+b and y=1+ax+b, because the contour lines for 10 peptides per pixel remain between these two lines.

To convert fractional mass to calculated mass defect, we used the following formula,

$$\text{Calculated Mass Defect} = \text{int}(aM+b-\text{int}(aM+b)-F) + \text{int}(aM+b)+1+F$$

where M stands for monoisotopic mass, F for fractional mass (that is, M-int(M)), a = 0.00050 and b= −0.59. This transformation merely adds an integer to the observed fractional mass. If points that lie below the dashed line in Fig3A are transformed upward by the integer n, then points in the same mass range that line above the dashed line would be transformed by moving upwards by n-1.

The calculation of Mass Defect from Fractional Mass is fairly accurate. Using the above formula, 98.8% of all peptides have the correct calculated Mass Defect value, the mismatches occurring at the edges of the data cloud where the addition of the correct integer value to the fractional mass is more prone to error. While the probability distribution map based on fractional mass shows a complex pattern (Fig 3B), the one based on calculated mass defect (Fig 3C) looks similar to the probability map based on true mass defect (Figure 3D). The probability distribution map derived from the calculated mass defect values matches 74% of the pixels of the map derived from the true Mass Defect. The mismatches occur mainly at masses> 4000 Da, where the mass defect data cloud expands beyond the two bounding lines mentioned above (Fig 3A). Below 4000 Da, 96% of pixels of the probability for mass<4000 are correct. Because the incorrect pixels are mostly at the edges of the data clouds where there are few peptides, 96.7% of all peptides and 99.9% of peptides with mass<4000 fall in the correct probability bins. Since most tryptic peptides in mass spectral data are observed with masses below this 4000 Da cut-off for MS and MS/MS analysis, the determination of empirical curves for phosphopeptide probability calculations were based below this cut-off.

Having generated a good approximation of the phosphorylation probability curve for the mass defect values, by using fractional masses, we examined the shape of the probability contour lines for several values of phosphorylation probability. For mass <4000 Da the probability contour lines for p=0.9, 0.8 and 0.5 are shown in Figure 3E. Each of these contour lines can be fitted by straight lines with regression parameters indicated in Table 1. Peptides having a mass and calculated mass deficit below one of these lines have a probability of phosphorylation that is greater than the indicated P value. At higher mass values the contour lines become more complex due to the scatter in the data and the “wrap around” of the fractional mass values. Therefore at mass values greater than 4000 Da, probability area need to be defined by two lines, one for the upper boundary and one for the lower one. Here, we will focus on the characterization of the different probability lines for peptides of mass < 4000 Da as a discriminator for phosphopeptides.

Among the various straight lines that one can use to discriminate between peptides having different probabilities of phosphorylation, we chose for further characterization the one whose parameters define the p=0.9 probability boundary (referred to as $L_{0.9}$ hereafter). By accounting for all of the peptides, we find that 95.4 % of peptides below this line have up to 3 phosphate-containing peptides, consistent with the definition of this line. We ascertained that this ratio is consistently valid throughout this area in the phosphorylation probability map. Specifically, we asked whether all phosphopeptides below the $L_{0.9}$ are truly within a local region of p>0.9 probability of phosphorylation. Our data was divided into bins of mass 10 Da wide and mass defect 0.01 Da high, so we categorized the peptides according to whether they were in a bin with phosphorylation probability ≥ or < 0.9, and whether they are below or above the line. The results are summarized in Table S3 in terms of sensitivity, specificity, positive predictive value and negative predictive value of our method. As a measure of the sensitivity, phosphopeptides within bins having a p value >0.9 are correctly positioned below $L_{0.9}$ about 98% of the time. Conversely, in terms of specificity, phosphopeptides belonging to a p<0.9 bin are correctly positioned above the line 97% of the time. The positive predictive value, that is, the fraction of phosphopeptides below the p>0.9 line belonging to a bin with a p>0.9 value, is 96%. The negative predictive value, that is, the fraction of phosphopeptides above the p>0.9 line belonging to a bin with a p <0.9, is 99%.

The distribution of potential peptides above and below the p>0.9 line for Human IPI database are shown in Table 2, along with the distributions for validated phosphopeptides. The fraction of valid P1 phosphopeptides (~16.5%) below the line is similar to that for Human IPI database P1 phosphopeptides (~15%), although the fractions for valid P2 and P3 ratios are about one third smaller than those for the potential peptides.

Next we considered whether the utility of our method for predicting phosphopeptides is affected by their mass measurement accuracy. We simulated experimental error in peptide mass determination by adding to the theoretical masses random noise of Gaussian distribution with a standard deviation of 1, 5, 20 and 100 ppm. We performed this randomization 50 times and each time, for each peptide in the database, we determined whether the corrupted mass falls below or above the $L_{0.9}$. As the error in the mass values increases, a greater proportion of peptides move across $L_{0.9}$ (Table S4A and S4B). However, up to 20 ppm error, peptides that move in one direction are replaced by a similar number of peptides of the same class, moving in the opposite direction, such that all measures of the classification accuracy (specificity, sensitivity, positive predictive value and negative predictive value) are affected by only a few percentage points (Tables S5A-S5D). As error increases from 0 to 20 ppm, Positive Predictive Value (the fraction of peptides below the line that are phosphorylated) decreases from 95.4% to 93.6 for all phosphopeptides, and from 80.6% to 76.9% for peptides with 1 phosphate. However the drop-off was more dramatic at 100 ppm with the fraction of peptides below the line that are phosphorylated decreasing from 95.4% to 76.4% for all phosphopeptides.

The analyses done up to this point were done on the database of tryptic peptides derived from the human proteome, containing zero to three phosphates. Probability calculations assumed that all peptide in this population are equally likely. A more accurate estimate of the probability of phosphorylation of peptides in a given bin of a plot of calculated mass defect vs. monoisotopic mass would be

$$P = (w_1N_1 + w_2N_2 + w_3N_3) / (w_0N_0 + w_2N_1 + w_2N_2 + w_3N_3)$$

where N_0, N_1, N_2 and N_3 refer to the number of peptide sequences in a bin, having 0, 1, 2, and 3 phosphates, respectively, and w_0-w_3 are weighing factors that reflect the likelihood of occurrence of these peptide classes in the peptide mixture being analyzed.

To estimate how our method would perform in an experimental setting where peptides of different phosphorylation states have different prevalence, we need to calculate probabilities using different w_0-w_3 factors. Our current state of knowledge regarding the distribution of phosphorylation sites in the human proteome is incomplete, but suggests that peptides with fewer phosphate groups are more likely to be observed, so the equi-probability assumption of phosphopeptide occurrence is unlikely to be valid in an experimental setting. On the other hand, the peptide distribution in a mixture that is being analyzed by tandem mass-spectroscopy may have been enriched for phosphopeptides, so the phosphorylation probability would need to reflect the proportion of phosphopeptide classes in that mixture. We attempted to quantitatively estimate the effect of varying the likelihood of occurrences of different phosphopeptide classes on the utility of the method we are presenting here.

The phospho.ELM database consists of proteins that have been experimentally validated to be phosphorylation substrates and whose phosphorylation sites is known [39]. By performing an *in-silico* tryptic digestion on the sequence of these proteins it can be seen that about 93% of peptide sequences from this small sample of the human proteome are unphosphorylated, while the proportion of peptides with 1, 2, and 3 peptides is about 5.3%, 1.4% and 0.4% (Table S1). The proportion of the human proteome that is phosphorylated is currently not known, although it is estimated from indirect evidence that up to a third of the mammalian proteome may be phosphorylated [40]. Based on this ratio, and assuming that the phospho.ELM database proteins

are a representative sample of the human proteome, we can adjust the likely proportion of tryptic peptides, having 0, 1, 2, or 3 phosphate groups to be 97.5%, 1.9%, 0.5% and 0.1%. We used these percentages as weighing factors w0-w3 in the probability formula given above and will refer to this peptide distribution as the "natural" distribution. We also considered the case where tryptic peptides derived from a natural distribution are ten-times enriched for phosphopeptides before MS analysis. In this case, the weighing factors become 0.80, 0.15, 0.04 and 0.01, respectively; we will refer to this distribution as the "10x enriched" distribution. Thus, we constructed probability distribution plots like Fig 3E for these two distributions, where the number of peptides within each bin was adjusted to reflect the peptide distribution case that was considered.

We used a Receiver Operating Characteristic response (ROC) curve to compare the discrimination ability of the phosphorylation probability plot for the "equi-probable", the "natural" and the "10x enriched" distribution cases as described in the previous paragraph. A ROC curve plots the True Positive Rate (TP/TP+FN) against the False Positive Rate (FP/FP +TN) as the discrimination threshold of the classifier function is varied (TP, true positive, FN, false negative, FP, false positive, TN, true negative). The shape of the curve reflects the tradeoff between the TPR and the FPR and summarizes the quality of the predictor function. A random predictor would give a 45 degree line, while a perfect predictor would have a single point at TPR=1, FPR=0. From the probability plots adjusted for the weighing factors of each phosphopeptide class, we constructed ROC curves from a series of p values ranging from 0 to 1. Thus, for each p value used to construct the ROC curve, phosphopeptides and unphosphorylated peptides in the bins having a phosphorylation probability greater than or equal to p were labeled as TP and FP, and peptides in those bins where this probability is less than p were labeled as FN and TN, then TPR and FPR values were calculated from the number of peptides in each category. As can be seen in Fig. 4, although the composition of the peptide mixture alters the shape of the ROC curve, the predictive power of the method remains strong even with a "natural" peptide mixture. The ratio of phosphopeptides to all peptides for the "equi-probable", "natural" and "10x enriched" distributions (for mass < 4000 Da) is 1.87, 0.019, 0.195, respectively, while below the $L_{0.9}$ line this ratio is 21.3, 0.18 and 1.72, respectively. Thus for the three cases, the enrichment for phosphopeptide that results by selecting peptides that are below the $L_{0.9}$ line is 11.4, 8.8 and 9.5-fold, respectively.

Lastly, we investigated the characteristics of the peptides falling below $L_{0.9}$ to see whether they have any distinguishing characteristics other than their phosphate groups. Because amino acids have different mass defect values, depending on their atomic composition, we examined how amino acid composition might affect the fraction of peptides falling below $L_{0.9}$.

While our algorithm for identifying peptides with a high probability of phosphorylation is based on the mass defect of the phosphate group, the mass defect of other atoms within amino acid residues influence the magnitude of the overall mass defect of each peptide. Thus, because sulfur and oxygen have mass defect values that are lower than carbon, sulfur- and oxygen-containing residues are more likely to be present below $L_{0.9}$. The strong correlation between the normalized mass defect of amino acids and their enrichment in peptides below $L_{0.9}$ confirms the principle by which our method works. Thus, the use of mass deficit to identify peptides having a high probability of phosphorylation also favors peptides having amino acids with low mass deficit.

Figure 5 shows number of average fractional composition of each amino acid (number of residues of a given amino acid divided by total number of residues in the peptide) for all human, unphosphorylated, tryptic peptides derived from the IPI database proteins. For each peptide, average fractional residue compositions were calculated for peptides above and below $L_{0.9}$ and with mass < 4000 Da. The large number of samples in the groups ensured that most comparisons

are different at a statistically significant level, although only a few are enriched or impoverished by a factor greater than 1.5-fold. Thus, non-phosphorylated peptides that fall below $L_{0.9}$ are enriched by about 4.2, 2.1, 1.7 and 1.6 times for Cys, Asp, Met, and Glu, while they are impoverished by 2.9, 2.6, 2.0 and 1.8 fold for Ile, Leu, Val and Lys, respectively. Similarly, validated peptides containing 1 phosphate that fall below the line are enriched about 2 fold for Cys, Met, Asp and Glu and impoverished about 1.6 to 2.1 fold for Ile, Leu and Val. The enrichment ratio of methionine is greater if it is oxidized. Note that the enrichment ratios for phosphopeptides is much larger than these ratios, as was mentioned above when comparing the three different cases of peptide compositions.

There was a high correlation between the mass-normalized mass defect (residue mass defect × mean residue mass / residue mass) of each residue and the logarithm of the enrichment ratio of that residue. The correlation was better for non-phosphorylated peptide (corr.=0.98) than for validated peptides having 1 or 2 phosphates (corr.=0.88; validated P3 peptides were not analyzed because of their low numbers; see Table S6). The slightly lower correlation coefficient found for the validated phosphopeptides is likely due to the enrichment of phosphorylated peptides for sequence motifs specific for phosphorylation substrates.

Although it is possible to estimate true phosphorylation probabilities as a function of mass for any given peptide mixture, one has to make a number of assumptions to model the distribution of phosphopeptides in that mixture. We attempted to do that for the “natural” and “10x enriched” distributions in order to show the general utility of our method. The $L_{0.9}$ line which was defined as the P=0.9 boundary in the “equi-probable” distribution, becomes like the P=0.1 boundary for the “natural” distribution and like the P=0.5 boundary for the “10x-enriched” distribution. Nevertheless, the fact that with each of these peptide distributions, peptides below the $L_{0.9}$ line are ~10 times more likely to be phosphorylated than all peptides points to the general applicability of using the mass defect of phosphopeptides to predict their phosphorylation probability. However, probability calculations performed on the “equi-probability” phosphopeptide distribution assumption, are of sufficient practical use that more complex computations seem unnecessary. A probability value calculated with the “equi-probability” assumption can serve as a gauge to “rank order” any given peptide on a scale of most likely being a phosphopeptide. Most tandem mass spectrometers make real time decisions on multiple precursor ions appearing simultaneously based on their charge state and intensity to proceed for MS/MS on a survey spectrum. The algorithms presented here would help make that decision on every surveyed spectrum based on the most likelihood of these peptides being a phosphopeptide. Clearly, not all theoretically predicted phosphopeptides exist in real biological systems resulting in different probability numbers. However, this would only result on the sliding scale of probabilities that would normally be set experimentally based on the ongoing success rate of triggering phosphopeptides. There may also be other outcomes on the probability calculations if other post-translationally modified peptides were included in the calculations. Molecules such as phospholipids and to lesser extent oxygen-rich glycopeptides would also have more negative mass defect compared to the unmodified peptides. However, the likelihood of having all these different class of peptides in the same experiment after specific isolation and enrichment protocols is lesser prior to mass spectrometric analysis.

One caveat of our method should be evident from our analysis of the amino acid composition of unphosphorylated peptides. Peptides with many oxygen and sulfur atoms have greater negative mass defect and therefore are more likely to occur below the $L_{0.9}$ line. Thus, oxygen- and/or sulfur-rich peptides will have a greater probability of being picked up by this method.

Conclusions

We have taken advantage of the fact that the mass defect value of the phosphate group will cause the total mass defect of a phosphopeptide to shift significantly away from unmodified peptides of the same nominal mass. We show here that with a simple two step approach this can be the basis for a method to identify phosphopeptides with relatively high probability from a survey mass spectrum. The first step is to calculate a peptide's mass defect from its experimental monoisotopic mass and the second is to determine whether the combination of mass and mass defect values for this peptide belong in a region of high phosphorylation probability. Regarding the mass defect calculation, we were able to define a function that performs correctly for 99% of all peptides. Regarding the assignment of a phosphorylation probability, we found that especially for peptides of mass < 4000, a single line suffices to delineate regions of high phosphorylation probability from those of lower probability. “Sliding scale” boundaries corresponding to other probability values can be selected similarly, as needed. We considered how likely a peptide's probability was to change if its mass had a measurement error associated with various instrumentation currently in use. We found that even with a measurement error of up to 20 ppm, the proportion of phosphopeptides above and below the selected probability line did not change significantly. This property can be very advantageous for the method to be applicable to wider range of tandem mass spectrometers with modest mass accuracies.

In order to assess and validate how generally applicable the mass defect approach is, we performed in-silico trypsin digestions on proteins from the IPI database and validated the results with the phospho.ELM database containing all possible phosphorylated proteins from the published literature. The set of published phosphopeptides derived from the phospho.ELM data set are similar to the set of theoretically derived phosphopeptides from IPI database in that both peptide sets show similar shifts in their mass deficits upon the addition of extra phosphate groups (Table 2) and the ratio of P1 peptides below and above $L_{0.9}$ are similar, at about 16%. More detailed analysis indicates that there are differences in the statistics of the regression lines in the Mass Defect vs. Monoisotopic Mass graph for the peptides from the two data sets (Table S2), in the distribution of phosphopeptides above and below $L_{0.9}$. Nevertheless, it is important to bear in mind that the validated peptide set contains known phosphopeptides, while the potential peptide set includes all phosphopeptides yet to be discovered along with those that are unlikely to be ever observed (e.g. at sites within membrane-spanning segments). Rather than trying to guess the unknown, all inclusive approach was taken in creating our set of potential phosphopeptides. The p=0.9 line $L_{0.9}$ that we characterized in this study would be a good guide that is a user selectable number and a tool in identifying new phosphopeptides during tandem mass spectrometric analysis of peptides.

There are possible future advantages to the utility of the mass defect principles associated with phosphopeptides. For example, the method presented in this paper could potentially aid in the database search of phosphoproteins by improving the scoring on selected peptides having a high probability of being phosphorylated. We have already generated a web site (http://phosphopeptide.med.yale.edu/) where the in-silico data used in this study can be queried with an interface that allows the user to enter monoisotopic, MH^+, masses and measurement errors for a group of peptides and obtain output that lists for each peptide the number of phosphorylated and unphosphorylated peptides within the given mass measurement error, and its probability of being phosphorylated. The user can drill down deeper into the output to determine the sequences of candidate peptides whose mass falls within the mass measurement error, along with annotation data regarding the source. Finally, the formulas and algorithm of our method can be easily incorporated into the system software/firmware of any real time data dependent decision making mass spectrometer, preferentially enabling phosphopeptides to be automatically picked for tandem mass spectrometric analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This project has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, under NIH contract No. N01-HV-28186 by NIH grant T15 LM07056 from the National Library of Medicine, and by National Institute on Drug Abuse, NIH grant 1 P30 DA018343-01

References

1. Cohen P. Nat Cell Biol 2002;4:E127–130. [PubMed: 11988757]
2. Hunter T. Cell 2000;100:113–127. [PubMed: 10647936]
3. Pawson T, Scott JD. Science 1997;278:2075–2080. [PubMed: 9405336]
4. Hubbard MJ, Cohen P. Trends Biochem Sci 1993;18:172–177. [PubMed: 8392229]
5. Wolschin F, Wienkoop S, Weckwerth W. Proteomics 2005;5:4389–4397. [PubMed: 16222723]
6. Gronborg M, Kristiansen TZ, Stensballe A, Andersen JS, Ohara O, Mann M, Jensen ON, Pandey A. Mol Cell Proteomics 2002;1:517–527. [PubMed: 12239280]
7. Beausoleil SA, Jedrychowski M, Schwartz D, Elias JE, Villen J, Li J, Cohn MA, Cantley LC, Gygi SP. Proc Natl Acad Sci U S A 2004;101:12130–12135. [PubMed: 15302935]
8. Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, Shabanowitz J, Hunt DF, White FM. Nat Biotechnol 2002;20:301–305. [PubMed: 11875433]
9. Larsen MR, Thingholm TE, Jensen ON, Roepstorff P, Jorgensen TJ. Mol Cell Proteomics 2005;4:873–886. [PubMed: 15858219]
10. McLachlin DT, Chait BT. Anal Chem 2003;75:6826–6836. [PubMed: 14670042]
11. Pinkse MW, Uitto PM, Hilhorst MJ, Ooms B, Heck AJ. Anal Chem 2004;76:3935–3943. [PubMed: 15253627]
12. Posewitz MC, Tempst P. Anal Chem 1999;71:2883–2892. [PubMed: 10424175]
13. Qian WJ, Goshe MB, Camp DG 2nd, Yu LR, Tang K, Smith RD. Anal Chem 2003;75:5441–5450. [PubMed: 14714534]
14. Tao WA, Wollscheid B, O'Brien R, Eng JK, Li XJ, Bodenmiller B, Watts JD, Hood L, Aebersold R. Nat Methods 2005;2:591–598. [PubMed: 16094384]
15. Zhou H, Watts JD, Aebersold R. Nat Biotechnol 2001;19:375–378. [PubMed: 11283598]
16. Annan RS, Huddleston MJ, Verma R, Deshaies RJ, Carr SA. Anal Chem 2001;73:393–404. [PubMed: 11217738]
17. DeGnore JP, Qin J. J Am Soc Mass Spectrom 1998;9:1175–1188. [PubMed: 9794085]
18. Le Blanc JC, Hager JW, Ilisiu AM, Hunter C, Zhong F, Chu I. Proteomics 2003;3:859–869. [PubMed: 12833509]
19. Schlosser A, Pipkorn R, Bossemeyer D, Lehmann WD. Anal Chem 2001;73:170–176. [PubMed: 11199962]
20. Schroeder MJ, Shabanowitz J, Schwartz JC, Hunt DF, Coon JJ. Anal Chem 2004;76:3590–3598. [PubMed: 15228329]
21. Steen H, Kuster B, Fernandez M, Pandey A, Mann M. Anal Chem 2001;73:1440–1448. [PubMed: 11321292]
22. Steen H, Mann M. Anal Chem 2002;74:6230–6236. [PubMed: 12510743]
23. Kendrick E. Anal. Chem 1963;35:2146–2154.
24. Hughey CA, Hendrickson CL, Rodgers RP, Marshall AG, Qian K. Anal Chem 2001;73:4676–4681. [PubMed: 11605846]
25. Kramer RW, Kujawinski EB, Hatcher PG. Environ Sci Technol 2004;38:3387–3395. [PubMed: 15260339]
26. Marshall AG, Rodgers RP. Acc Chem Res 2004;37:53–59. [PubMed: 14730994]
27. Wu Z, Rodgers RP, Marshall AG. Anal Chem 2004;76:2511–2516. [PubMed: 15117191]

28. Gay S, Binz PA, Hochstrasser DF, Appel RD. Electrophoresis 1999;20:3527–3534. [PubMed: 10612279]
29. Conrads TP, Anderson GA, Veenstra TD, Pasa-Tolic L, Smith RD. Anal Chem 2000;72:3349–3354. [PubMed: 10939410]
30. Boltz, SA.; Brown, M.; Niehauser, S.; Phillips, R.; Wolff, J.; Amster, JI. Proceedings of the 51st ASMS Conference on Mass Spectrometry and Allied Topics; Montreal, Quebec, Canada. 2003.
31. Hernández, H.; Vichchulada, P.; Niehauser, S.; Amster, J. Nashville, TN; 2004.
32. Hall MP, Ashrafi S, Obegi I, Petesch R, Peterson JN, Schneider LV. J Mass Spectrom 2003;38:809–816. [PubMed: 12938101]
33. Hall, MP.; Ashrafi, S.; Petesch, R.; Obegi, I.; Schneider, LV. Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics, Orlando, FL; 2002.
34. Whetstone PA, Butlin NG, Corneillie TM, Meares CF. Bioconjug Chem 2004;15:3–6. [PubMed: 14733576]
35. Schneider LV, Hall MP. Drug Discov Today 2005;10:353–363. [PubMed: 15749284]
36. Jones JJ, Stump MJ, Fleming RC, Lay JO Jr. Wilkins CL. J Am Soc Mass Spectrom 2004;15:1665–1674. [PubMed: 15519235]
37. Zhang H, Zhang D, Ray K. J Mass Spectrom 2003;38:1110–1112. [PubMed: 14595861]
38. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. Proteomics 2004;4:1985–1988. [PubMed: 15221759]
39. Diella F, Cameron S, Gemund C, Linding R, Via A, Kuster B, Sicheritz-Ponten T, Blom N, Gibson TJ. BMC Bioinformatics 2004;5:79. [PubMed: 15212693]
40. Johnson SA, Hunter T. Nat Methods 2005;2:17–25. [PubMed: 15789031]

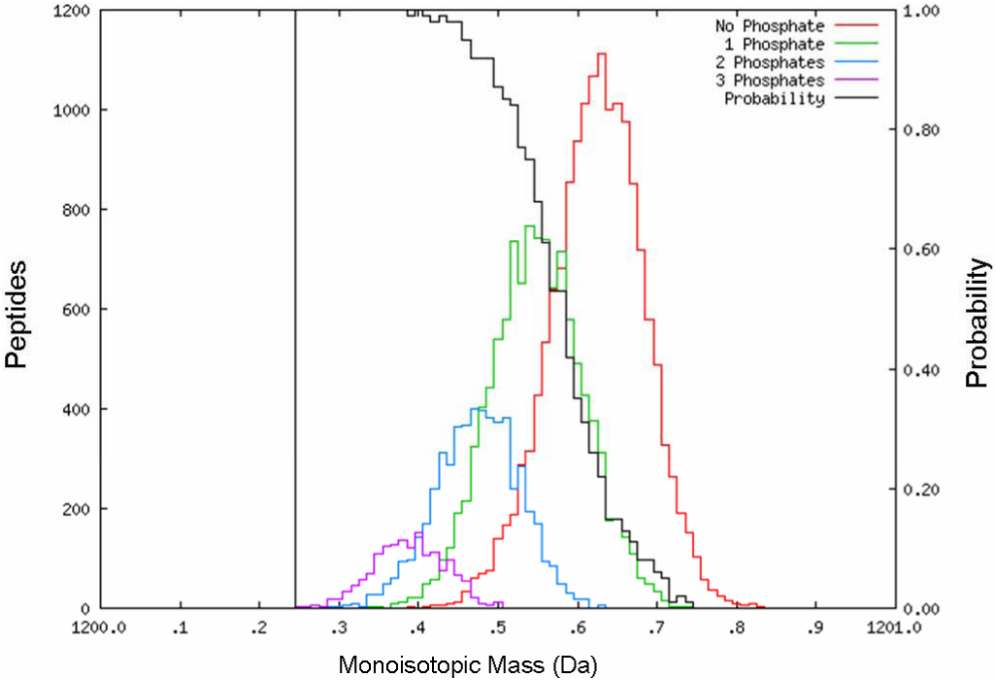

Figure 1.

Distribution of peptide numbers and phosphorylation probability for tryptic peptides with 0, 1, 2, and 3 phosphates, and, in the mass range 1200 to 1201. Probability of phosphorylation is (N1+N2+N3)/(N0+N1+N2+N3), where N0, N1, N2 and N3 are the number of peptides with 0, 1, 2, and 3 phosphates, respectively, within each bin of the histogram. Each bin size is 0.01 Da.

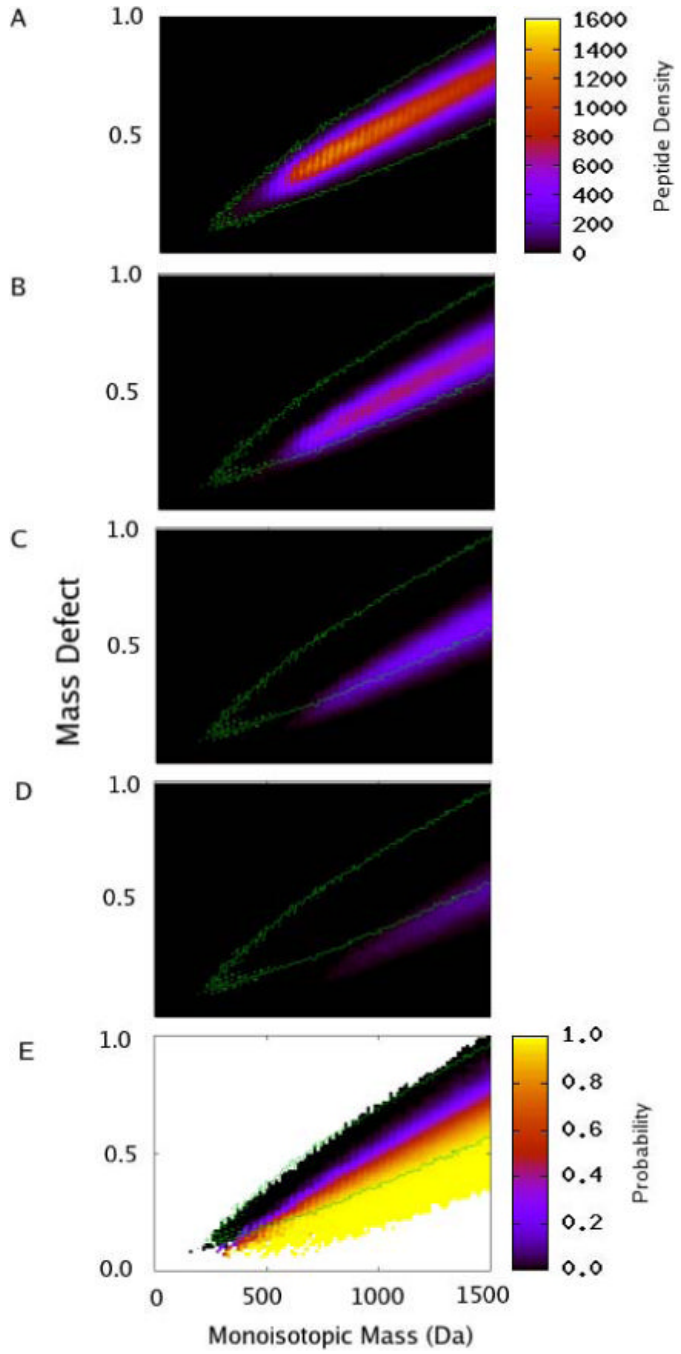

Figure 2.

Density map of peptide numbers and phosphorylation probability for tryptic peptides with 0, 1, 2 and 3 phosphates. **A-D,** Density map of peptides with 0, 1, 2 and 3 phosphates respectively, as a function of Mass Defect and Mass. Colors denote number of peptide within one pixel of the map, each pixel having dimension of 10 Da Mass × 0.01 Da Mass Defect. Color map in panel A applies to panels B-D as well. Green contour in Panel A shows boundary or region having ≥10 peptides per pixel and is reproduced in panels B-E for reference. **E**, Phosphorylation probability distribution as a function of Mass Defect and Mass. Probability of phosphorylation is (N1+N2+N3)/(N0+N1+N2+N3), where N0, N1, N2 and N3 are the number of peptides with 0, 1, 2, and 3 phosphates, respectively, within each pixel.

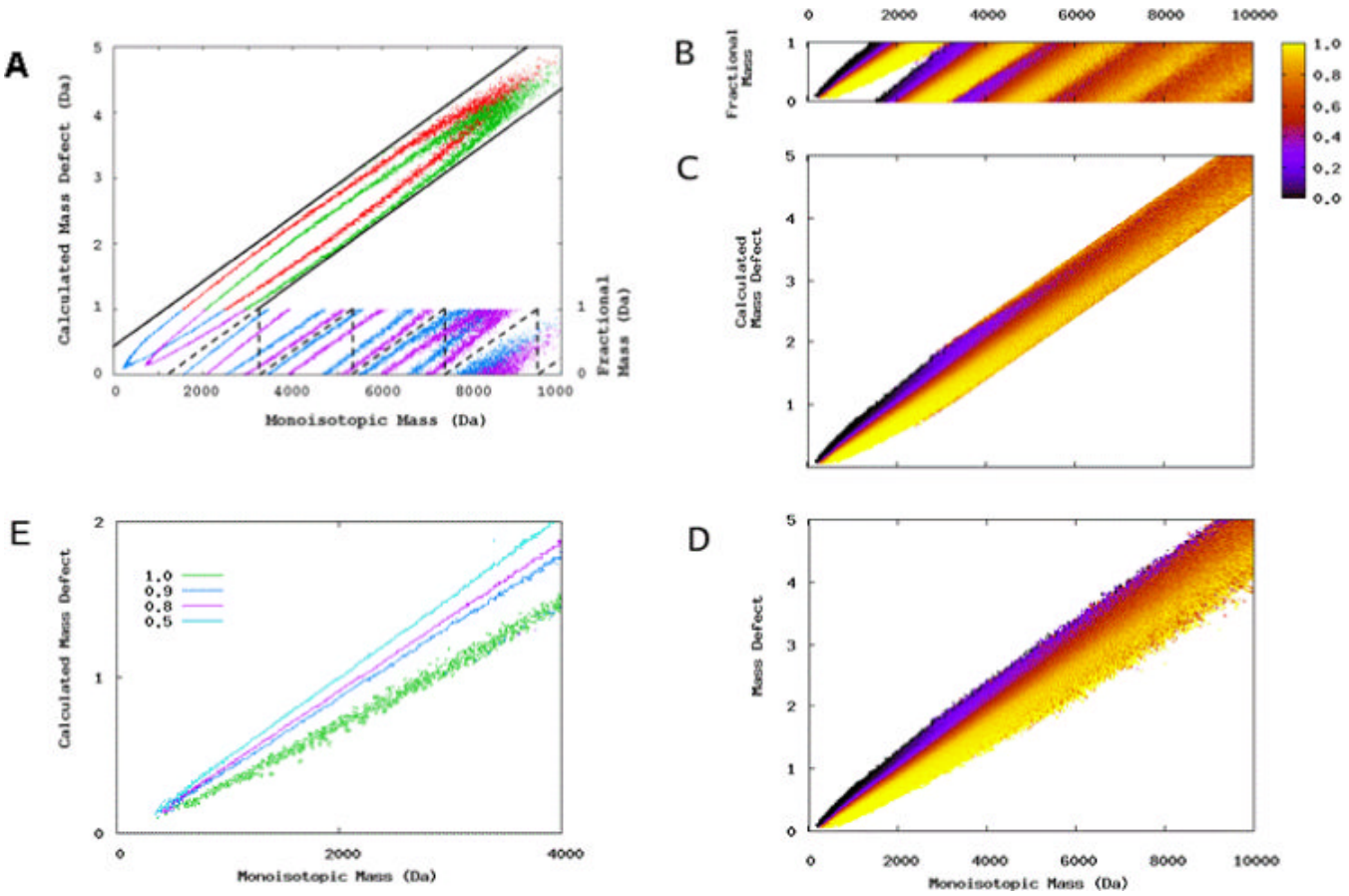

Figure 3.

A. Density Map for peptides with 0 and 3 phosphates, shown as a contour lines (in red and green, respectively) for regions having ≥10 peptides per pixel, as a function of Mass and Fractional Mass (Right Axis; blue and purple lines) and of Mass and Calculated Mass Defect (Left Axis, red and green lines). By converting Fractional Mass values to calculated Mass Defect (see text for details), the discontinuous contour lines in blue and purple are transformed into the continuous contour lines in red and green, respectively. Where the blue and red curves, or green and purple curves coincide, the blue and purple curves, respectively, are plotted only. The Fractional Mass data points between 0 and 1 are translated to equivalent positions between the two parallel solid lines that are spaced 1.0 Da apart. The dashed zigzag line corresponds to the function aM+b − int(aM+b) (see text for details). **B**. Phosphorylation probability distribution map as a function of Monoisotopic Mass and Fractional Mass. Color Map shows probability of phosphorylation, as defined in text, for each pixel of 10 Da Mass × 0.01 Da Mass Defect. C. Phosphorylation probability distribution map as a function of Monoisotopic Mass and Calculated Mass Defect. **D**. Phosphorylation probability distribution map as a function of Monoisotopic Mass and true Mass Defect. **E**. Probability distribution data in panel C re-plotted as a contour map to show linearity of selected contour lines. Note different scale of the graph axes.

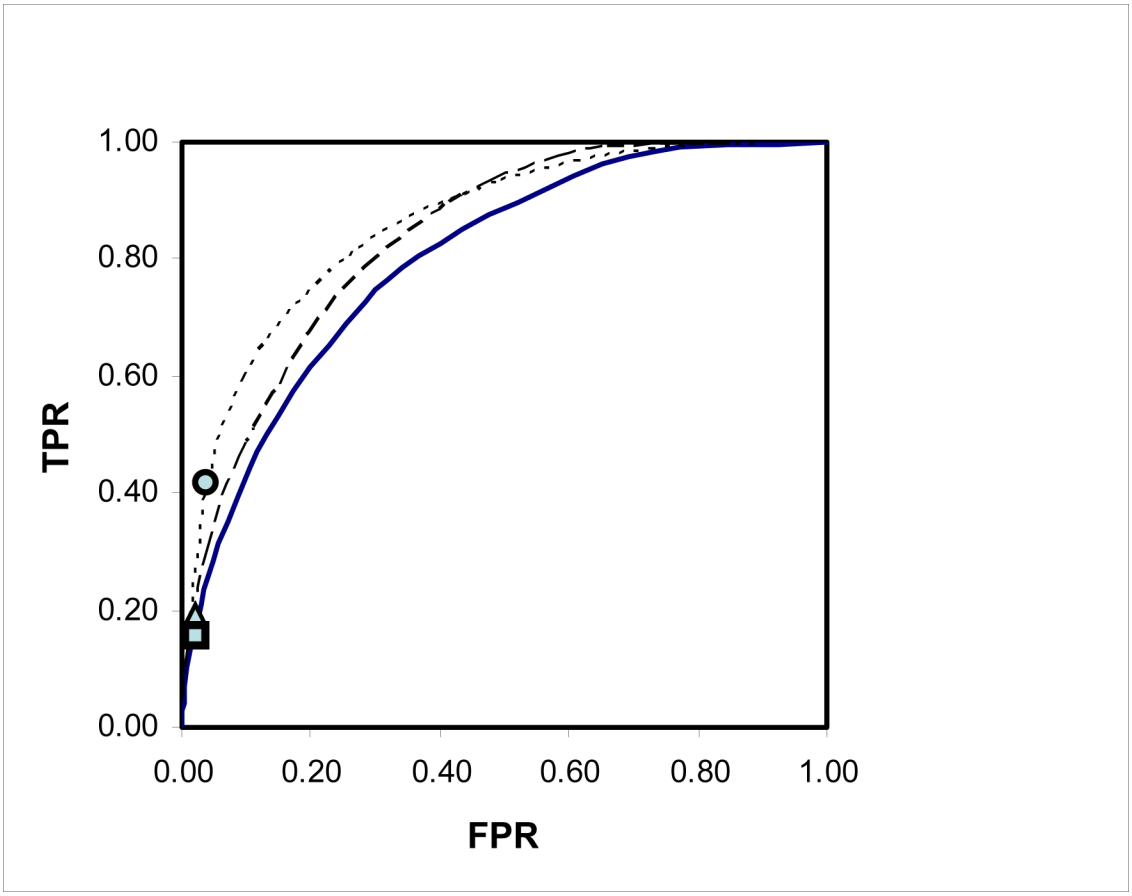

Figure 4.
Receiver Operating Characteristic (ROC) curve of phosphopeptide prediction for three theoretical peptide mixtures, the “equi-probability” distribution (dotted line), “natural” distribution (dashed line), “10x enriched” distribution (continuous line); see text for details. TPR, true positive rate, FPR, false positive rate. The TPR and FPR values for the $L_{0.9}$ line (defined in the text) is indicated on these three curves with a circle, triangle and a square.

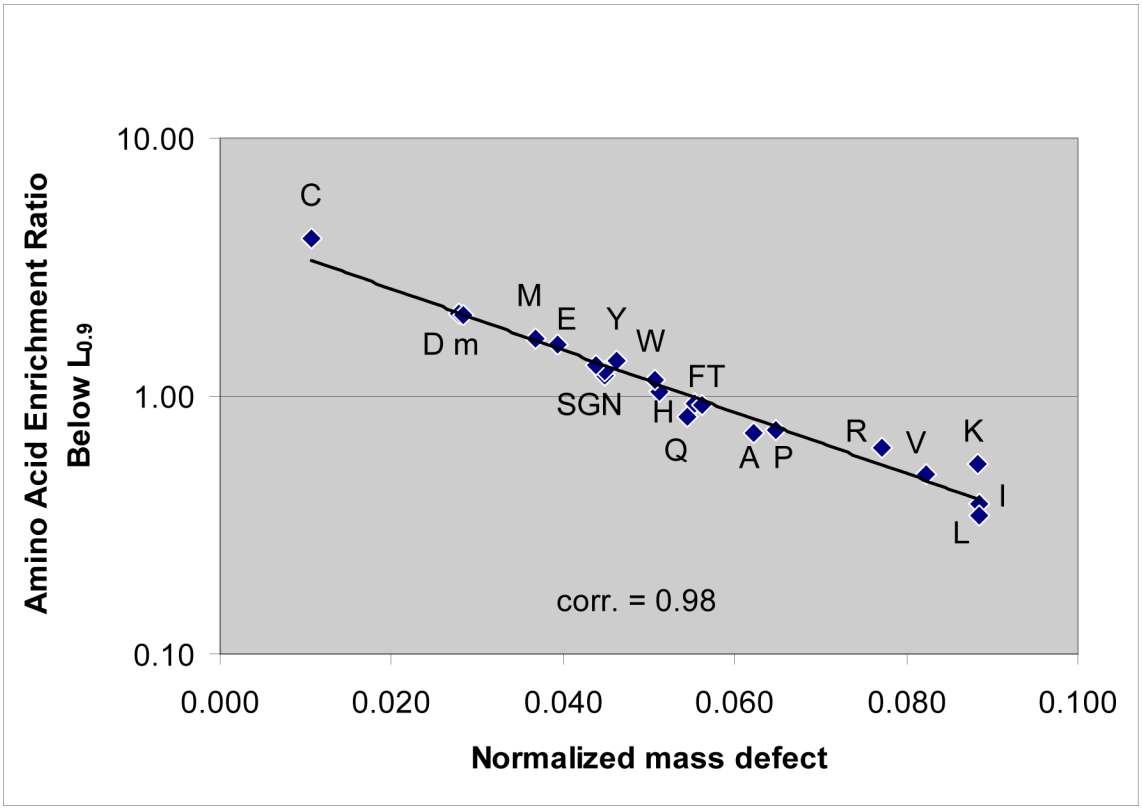

Figure 5.

Enrichment of amino acids with low mass defect below $L_{0.9}$. For each amino acid, the ratio of its composition for unphosphorylated peptides above and below the p=0.9 line was determined (Table S6) and plotted against the mass defect of that amino acid, normalized for its mass. Standard one-letter amino acid codes are placed next to each data point, except for **m**, which stands here for methionine sulfoxide.

Table 1

Linear regression parameters for phosphorylation probability boundary lines for peptides of mass<4000 Da.

P	a	b
0.90	0.000457±0.0000006	−0.0448±0.0029
0.80	0.000485±0.0000005	−0.0901 ± 0.0043
0.50	0.000519±0.0000005	−0.0438 ± 0.0061

Table 2

Percentage of peptides below $L_{0.9}$, for peptides derived from the IPI database and for validated phosphopeptides.

		P0	P1	P2	P3
Human IPI database	All peptides:	3,216,923	2,698,158	1,983,717	1,335,576
	Below $L_{0.9}$ (%)	114,450 (3.6%)	480,806 (17.8%)	965,217 (48.7%)	1,010,395 (75.7%)
phospho.ELM database, human subset	All peptides:	NA	12,211	3,185	885
	Below $L_{0.9}$ (%)	NA	2,019 (16.5 %)	1,076 (33.8%)	358 (40.5%)