Published in final edited form as:

J Proteome Res. 2008 June ; 7(6): 2195–2203. doi:10.1021/pr070510t.

Linear Discriminant Analysis-Based Estimation of the False Discovery Rate for Phosphopeptide Identifications

Xiuxia Du¹, Feng Yang¹, Nathan P. Manes¹, David L. Stenoien¹, Matthew E. Monroe¹, Joshua N. Adkins¹, David J. States², Samuel O. Purvine¹, David G. Camp II¹, and Richard D. Smith^{1,*}

¹Fundamental and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA 99352

²University of Michigan, School of Medicine, Ann Arbor, MI 48109

Abstract

The development of liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) has made it possible to measure phosphopeptides on an increasingly large-scale and high-throughput fashion. However, extracting confident phosphopeptide identifications from the resulting large dataset in a similar high-throughput fashion remains difficult, as does rigorously estimating the false discovery rate (*FDR*) of a set of phosphopeptide identifications. This article describes a data analysis pipeline designed to address these issues. The first step is to re-analyze phosphopeptide identifications that contain ambiguous assignments for the incorporated phosphate(s) to determine the most likely arrangement of the phosphate(s). The next step is to employ an expectation maximization algorithm to estimate the joint distribution of the SEQUEST scores. A linear discriminant analysis is then performed to determine how to optimally combine peptide scores (in this case, SEQUEST) into a discriminant score that possesses the maximum discriminating power. Based on this discriminant score, the p- and q-values for each phosphopeptide identification are calculated, and the phosphopeptide identification *FDR* is then estimated. This data analysis approach was applied to data from a study of irradiated human skin fibroblasts to provide a robust estimate of *FDR* for phosphopeptides, and has been coded into a software package that is freely available (http://ncrr.pnl.gov/downloads/data/Du2008_Supplementary_Data.zip).

Keywords

False Discovery Rate; phosphoproteomics; expectation maximization; linear discriminant analysis; p-value; q-value; Bayesian analysis

Introduction

Post-translational modifications of proteins have critical physiological roles in biological systems, notably enabling signal transduction and protein activation and inhibition.¹ As a result, identification of modified proteins is an extremely important component of studies aimed at determining the mechanisms of biological function and malfunction. Phosphorylation is one of the most important of the post-translational modifications, and is required for many cellular processes such as cell cycle transition, differentiation, and regulated proteolysis.^{2, 3} The development of liquid chromatography coupled with tandem mass spectrometry (LC-MS/

*To whom correspondence should be addressed. Richard D. Smith., Dr. Richard D. Smith, P.O. Box 99 / K8-98, Richland, WA 99354, Email. lds@pnl.gov

MS)⁴ has made it possible to analyze phosphopeptides on an increasingly large-scale and high-throughput fashion. Hundreds to thousands of phosphopeptides can be measured in a single bottom-up LC-MS/MS analysis; however, extracting confident phosphopeptide identifications from the resulting large dataset in a high-throughput fashion remains difficult, as does rigorously estimating the false discovery rate (*FDR*) for a set of phosphopeptide identifications.

A number of statistical approaches have been applied to assess the quality of peptide identifications.⁵⁻⁹ Among these approaches are the Bayesian method used by PeptideProphet⁸ that assigns a probability value to each peptide identification and the target-decoy search strategy⁹ that is used to estimate the *FDR*. Both of these approaches use results generated by peptide search engines such as SEQUEST¹⁰ and X! Tandem.¹¹ For example, in PeptideProphet SEQUEST scores for *XCorr*, ΔCn , *Sp*, and *RankSp* are linearly combined to calculate a discriminant score before Bayesian theory is applied to estimate peptide identification probability scores. The weighting factors employed to combine the SEQUEST output scores into a discriminant score are obtained by static data training (i.e. from prior training datasets), and are fixed for all datasets regardless of biological sample and the type of mass spectrometer. However, the data quality of MS/MS spectra varies greatly for different MS instruments and biological samples and the results of peptide search engines generally differ as well. Therefore, the weighting factors should be MS/MS dataset-dependent, and dynamic training (i.e. training from the dataset of concern) should be performed to obtain factors specific to each dataset.

In a target-decoy analysis, the MS/MS spectra are searched against forward and reversed protein sequences, and the resulting identifications are used to estimate a *FDR* (which is equal to twice the number of reversed peptide identifications divided by the total number of peptide identifications).¹² This approach is based on the assumption that the distribution of the peptide identification scores (e.g., *XCorr*, ΔCn) of the incorrectly identified peptides from the forward database search is the same as the corresponding distribution from a reversed sequence search. There are two general approaches for performing target-decoy searches. One is to perform two separate searches against the target and decoy databases and the other is to perform a single search against a concatenated target plus decoy database. In the former approach, the computed *FDR* tends to be a conservative overestimate.⁵ In the latter approach, there are more candidates for each MS/MS spectrum, and thus ΔCn is reduced. This reduction reduces the discriminative power of ΔCn to distinguish among correctly and incorrectly identified peptides (see supplementary material). Regardless of which approach is used, a target-decoy analysis requires a longer search time than a target-only database search.

When applied to analyze phosphopeptides identified by SEQUEST, both the PeptideProphet and target-decoy approaches fail to take into account a specific issue related to the phosphorylation site assignment and to ΔCn . For each MS/MS spectrum, SEQUEST outputs ten peptides ranked by *XCorr*. If a phosphopeptide contains multiple potential phosphorylation sites and contains fewer phosphorylated residues than potential phosphorylation sites, then the *XCorr* value of the correct phosphopeptide identification will often be comparable to that of the same phosphopeptide (but with its incorporated phosphates rearranged). This occurrence results in a very small ΔCn value, and the “top hit” is not necessarily the correct identification. Therefore, two scores are needed for each phosphopeptide identification: 1) a score indicative of the confidence in the phosphorylation site assignment(s) and 2) a ΔCn variant related to the confidence in the identification of the amino acid sequence of the phosphopeptide.

In this study, we addressed the aforementioned issues by applying a data analysis pipeline that processed phosphopeptide search results and rigorously estimated the *FDR*. For each dataset, a dynamic data training routine is performed that calculates dataset-dependent weights designed to optimally combine the peptide scores (specifically SEQUEST) into a discriminant

score. The *p*- and *q*-values for each identified phosphopeptide are then calculated. The data training is charge state dependent, and is accomplished through an expectation maximization (EM) algorithm and a linear discriminant analysis (LDA). Lastly, a single, composite FDR is calculated for the entire set of filter-passing phosphopeptide identifications (i.e., this set includes identifications with different parent-ion charge states). This data analysis pipeline was initially applied to data from a study of irradiated human fibroblasts, and has been made into a software package that is freely available.

Methods

The *FDR* estimation algorithm was applied to phosphoproteome datasets from a study of irradiated human skin fibroblasts (see the supplemental material for more experimental detail). Briefly, protein samples were digested and methylated, and then phosphopeptides were enriched using a Fe^{3+} immobilized metal affinity chromatography (IMAC) protocol.^{13, 14} After IMAC enrichment, the sample was analyzed by LC-MS/MS using a ThermoElectron LTQ-Orbitrap. The data analysis procedure consisted of a sequential series of steps that began with SEQUEST analyses of the MS/MS spectra and concluded by estimating the *FDR* of the entire set of filter-passing phosphopeptide identifications (Figure 1 is a flow chart of the data analysis pipeline).

SEQUEST analysis, peptide filtering, and re-assignment of phosphorylation site(s)

MS/MS spectra were searched using SEQUEST (Sequest Cluster version 27 revision 12 from Bioworks 3.2, Thermo Electro Corp., Waltham, MA) against the Human International Protein Index database¹⁵ (version 3.20, 61,225 protein sequences, www.ebi.ac.uk/IPI, European Bioinformatics Institute, Cambridge, UK). The search parameters were:

1. Fully tryptic peptide termini (allowing ≤ 2 missed cleavages) (amino- and carboxy-termini were considered tryptic termini).
2. Dynamic modifications: an addition of 79.9663 Da to serine, threonine, and tyrosine residues (phosphorylation).
3. Static modifications: an addition of 14.0157 Da to aspartic acid, glutamic acid, and the carboxy-terminus (methylation is performed to improve IMAC enrichment).
4. Precursor ion mass tolerance: 0.05 Da
5. Fragment ion mass tolerance: 0.5 Da (*m/z*)
6. Maximum number of the same amino acid that can be dynamically modified in a phosphopeptide: 3

SEQUEST generally outputs 10 peptide sequences for each MS/MS spectrum that are ranked by *XCorr*, and the peptide with the highest *XCorr* is considered the “top hit” (or “1st hit”). Only MS/MS spectra that result in phosphopeptide top hits are kept for further analysis. This filtering step is performed because during collision-induced dissociation in the ion trap, unmodified peptides are usually fragmented more efficiently than phosphopeptides and often have higher *XCorr* and ΔCn scores. As a result, phosphopeptide SEQUEST scores are generally lower than the corresponding unmodified peptide scores. Therefore, the unmodified and modified peptide identifications should not be analyzed together; otherwise many correct phosphopeptides would likely be filtered away resulting in a high false negative rate. However, this filtration step might result in some false negative identifications in a different way which can happen when an MS/MS spectrum has a first hit that is an unmodified peptide and a second (or even third) hit that is a phosphopeptide and has almost as large an *XCorr* score. In this scenario, it can be difficult to tell in an automated fashion which hit is more confident, the unmodified,

first-hit or the phosphorylated, second-hit (or even the third-hit). Thus, it is still preferred to filter to remove spectra that result in top hits that are unmodified.

If an identified phosphopeptide contains a larger number of potential phosphorylation sites than incorporated phosphates, then there are multiple possible arrangements of the incorporated phosphates. Generally, this set of phosphopeptides has similar theoretical MS/MS spectra, and thus the *XCorr* values will be comparable. As a result, these closely related phosphopeptides will generally be among the ten ranked hits that SEQUEST assigns to the MS/MS spectrum. Since their *XCorr* values do not differ significantly from one another, the top hit is not necessarily the highest confidence identification. Therefore, a more careful analysis of the phosphopeptide variants is needed to identify the most confident identification. This analysis is performed by calculating the peptide score¹⁶ for each variant of the top hit and the phosphopeptide identification with the highest peptide score is considered the “true top hit”.

Calculation of $\Delta Cn'$

ΔCn is defined as the normalized difference between the *XCorr* of the peptide identification and the *XCorr* of the succeeding peptide identification (by *XCorr* rank), i.e.

$$\Delta Cn(n) = \frac{XCorr(n) - XCorr(n+1)}{XCorr(n)} \quad (1)$$

where n is the peptide identification *XCorr* rank. Therefore, ΔCn of the top hit is the normalized difference between the *XCorr* of the first and second hit. However, after the true top hit is obtained by rearranging the incorporated phosphates of the original top hit, ΔCn of the original top hit is unrelated to the confidence of the identification of the amino acid sequence of the phosphopeptide, and a ΔCn variant, $\Delta Cn'$, needs to be computed. Obtaining $\Delta Cn'$ involves determining the *XCorr* of the true top hit, and then searching for the “true second hit”. The true second hit is the highest-ranked identification that has a different amino acid sequence than the true top hit.

In general, the true top hit is among the ten hits, and its *XCorr* value is used. In the rare case that the true top hit is not among the ten hits, the peptide identification having the same amino acid sequence and having the lowest *XCorr* value is used. The *Sp* scores of the true top hit were determined similarly.

Occasionally, all ten hits have the same amino acid sequence and the only difference among them is the arrangement of the incorporated phosphates. Since there is not a peptide identification that satisfies the requirement to be a true second hit, these spectra are excluded from the data analysis.

Transformation of *XCorr*, *Sp*, and $\Delta Cn'$

The confidence of each true top hit is quantified using three parameters: *XCorr*, *Sp*, and $\Delta Cn'$. *XCorr* reflects the strength of the cross correlation between the experimental and theoretical MS/MS spectra, and longer peptide sequences tend to result in a larger *XCorr* values. To compare quantified confidences of peptide identifications with differing peptide lengths, *XCorr* needs to be normalized against the phosphopeptide length so that this bias is removed. The normalization equation used by PeptideProphet⁸ is used here:

$$\overline{XCorr} = \frac{\ln(XCorr)}{\ln(L)} \quad (2)$$

where L is the phosphopeptide length.

To make the joint distribution of $XCorr$, Sp , and $\Delta Cn'$ approximately normal, Sp and $\Delta Cn'$ are also transformed:

$$\overline{Sp} = \frac{\ln(Sp)}{10} \quad (3)$$

$$\overline{\Delta Cn'} = \sqrt{\Delta Cn'} \quad (4)$$

where the transformation of $\Delta Cn'$ is the same as that reported by Lopez-Ferrer, et al.¹⁷

Estimation of the mixed joint distribution using EM

For each of the three SEQUEST scores, a higher value is indicative of a higher likelihood that a peptide identification is correct. Therefore, the correctly and incorrectly identified phosphopeptides form two different clusters in $\overline{XCorr} - \overline{Sp} - \overline{\Delta Cn'}$ space. The joint probability density function (pdf) of these three values can be utilized to quantify the confidence of the phosphopeptide identifications, and this pdf can be estimated using EM. EM is a method of calculating maximum-likelihood estimates of parameters of an underlying distribution from a statistical sample.¹⁸⁻²⁰

Assuming that the joint distributions of the two clusters are both normal, the observed $\overline{XCorr} - \overline{Sp} - \overline{\Delta Cn'}$ triplet can be approximated by a Gaussian Mixture Model (GMM) and the mixed density function is:

$$p(X) = \pi_0 p_0 + \pi_1 p_1 \\ = \pi_0 \frac{1}{\sqrt{2\pi}|\Sigma_0|} e^{-\frac{1}{2}(X-\mu_0)^T \Sigma_0^{-1} (X-\mu_0)} + \pi_1 \frac{1}{\sqrt{2\pi}|\Sigma_1|} e^{-\frac{1}{2}(X-\mu_1)^T \Sigma_1^{-1} (X-\mu_1)} \quad (4)$$

where p_0 and p_1 denote the pdf of the incorrect and correct phosphopeptide identifications, respectively. π_0 and π_1 are the proportions of the identifications belonging to p_0 and p_1 , respectively, and they satisfy

$$\pi_0 + \pi_1 = 1 \quad (5)$$

X is a column vector:

$$X = \begin{bmatrix} \overline{XCorr} \\ \overline{Sp} \\ \overline{\Delta Cn'} \end{bmatrix} \quad (6)$$

μ_0 and μ_1 are the means of p_0 and p_1 , respectively. Σ_0 and Σ_1 are the covariance matrices of p_0 and p_1 , respectively. EM is used to estimate $\pi_0, \pi_1, \mu_0, \mu_1, \Sigma_0$, and Σ_1 .

Let $[D(i, j)], i = 0, \dots, N-1, j = 0, \dots, 2$ be a matrix of the SEQUEST scores of all of the identifications, where N is the total number of identifications, and the 1st, 2nd, and 3rd column of $[D(i, j)]$ are the $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$ values, respectively (i.e., each row of $[D(i, j)]$ corresponds to a single identification). EM consists of iterations of two sequential steps: expectation and maximization. During each iteration, the following two calculations are performed:

1. Expectation:

$$g(i, k) = \frac{\pi_k p_k(x_i)}{\pi_0 p_0 + \pi_1 p_1} \quad (7)$$

where $[g(i, k)]$ is a matrix with $i = 0, \dots, N-1, k = 0, 1$. x_i is the i^{th} row in data matrix $[D(i, j)]$.

2. Maximization:

$$\pi_k = \frac{\sum_{i=0}^{N-1} g(i, k)}{N} \quad (8)$$

$$\mu_k = \frac{\sum_{i=0}^{N-1} x_i * g(i, k)}{\sum_{i=0}^{N-1} g(i, k)} \quad (9)$$

$$\Sigma_k = \frac{\sum_{i=0}^{N-1} (x_i - \mu_k)^T (x_i - \mu_k) g(i, k)}{\sum_{i=0}^{N-1} g(i, k)} \quad (10)$$

This iteration continues until the estimated parameters converge.

Since EM is an iterative process, the initial values of $\pi_0, \pi_1, \mu_0, \mu_1, \Sigma_0$, and Σ_1 have to be provided. This is achieved through a preliminary clustering of the phosphopeptide identifications. All of the identifications are clustered into one of two clusters, cluster 0 and 1 that correspond to incorrectly and correctly identified peptides, respectively. This clustering consists of two sequential steps. During the first step, the difference ΔM between the measured and theoretical precursor masses (in ppm) is calculated and the distribution of ΔM is plotted. Generally, the distribution consists of a central dense region and a background as illustrated in Figure 5. The maximum lower and minimum upper boundaries of ΔM are then identified, and between them is the central dense region. The boundaries of the central dense region are identified by visual inspection. These boundaries can also be identified by an automatic processing of the histogram. Those peptides that have ΔM values that fall outside of the central region are more likely to be incorrect identifications, and are placed into cluster 0. The remaining peptides are placed into either cluster 0 or cluster 1 by k-means clustering using $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$. With all of the peptide identifications clustered, the initial values of $\pi_0, \pi_1, \mu_0, \mu_1, \Sigma_0$, and Σ_1 can be calculated as the sample statistics of these two clusters.

Linear discriminant analysis

After the pdf in Eqn. (4) is estimated, $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$ need to be combined into a single discriminant score for each peptide identification. Combining them so that the resultant discriminant score has the most discriminative power to distinguish correct and incorrect peptide identifications can be achieved using a linear discriminant analysis (LDA).²¹

To perform LDA, the membership of each peptide identification has to be determined first, which is achieved through a maximum likelihood analysis. For each identification, the probability of the identification belonging to the distributions p_0 and p_1 are calculated as π_0 and π_1 , respectively. Whichever is larger determines the membership of the peptide identification.

LDA calculates the weights used to linearly combine $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$ into a discriminant score. These weights are the components of the eigenvector corresponding to the largest eigenvalue of the following matrix:

$$A = W^{-1}B \quad (11)$$

where

$$W = \frac{1}{n_0 + n_1} (n_0 * \text{cov}_0 + n_1 * \text{cov}_1) \quad (12)$$

$$B = (\text{avg}_0 - \text{avg}) * (\text{avg}_0 - \text{avg})^T + (\text{avg}_1 - \text{avg}) * (\text{avg}_1 - \text{avg})^T \quad (13)$$

where n_0 and n_1 are the number of identifications in each cluster, avg_0 and avg_1 are column vectors of the sample means of $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$ in each cluster, and cov_0 and cov_1 are the covariance matrices of $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$ in each cluster, and avg is a column vector of the sample means of $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$ of all of the identifications from both clusters.

The eigenvector corresponding to the largest eigenvalue of A is $U = [u_1, u_2, u_3]^T$. The discriminant score for each peptide identification can then be calculated:

$$F_i = u_1 * \overline{XCorr}_i + u_2 * \overline{Sp}_i + u_3 * \overline{\Delta Cn'}_i, \quad i = 0, \dots, N - 1 \quad (14)$$

Estimation of the discriminant score pdf using EM

The discriminant scores obtained using Eqn. (14) are samples of F , the discriminant score. The pdf of F will again be modeled as a mixture of two components that correspond to incorrectly and correctly identified peptides. Assuming that F can be described by a Gaussian Mixture Model, the parameters of this pdf can be estimated by performing a second EM, similar to the one that was used to estimate the joint distributions of $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$.

FDR estimation

The discriminant score calculated above is used as the statistic to estimate the *FDR* of the set of filter-passing identifications. *FDR* estimation involves testing multiple hypotheses with each peptide identification corresponding to a single hypothesis test. Table 1 illustrates the possible outcomes when a significance threshold is applied to all of the phosphopeptide identifications. The null hypothesis is the hypothesis that the identification is incorrect and the alternative hypothesis is the hypothesis that the identification is correct. The terms “declared non-significant” and “declared significant” indicate non-filter-passing and filter-passing identifications, respectively. R is the total number of filter-passing identifications. V is the number of false positive filter-passing identifications. S is the number of true positive filter-passing phosphopeptide identifications. N is the total number of identifications. V , R , and S are random variables. It is common statistical practice to write the overall error measure in terms of an expected value and the *FDR* is thus defined as:

$$FDR = E\left(\frac{V}{R} | R > 0\right) \Pr(R > 0) \quad (15)$$

where E is the expectation operator. Because m is usually very large in proteome-wide studies, $\Pr(R > 0) \approx 1$, the *FDR* can be estimated using the following equation:

$$FDR \approx E\left(\frac{V}{R} | R > 0\right) \approx \frac{E(V)}{E(R)} = \frac{E(V)}{E(V) + E(S)} \quad (16)$$

Just as the false positive rate (*FPR*) is associated with a p-value from a single hypothesis test, the *FDR* is associated with q-values from the multiple hypotheses test. The q-value of each identification can be described as the expected proportion of false positive identifications among all identifications that are determined to be either as confident or more confident than the one under consideration, and therefore it quantifies the significance of each identification. Calculating the q-value for each identification and thresholding the identifications at a q-value of λ ($0 \leq \lambda \leq 1$) produces a set of identifications with a proportion of at most λ expected to be false positives. Therefore, these q-values can be used to filter the data and achieve a desired

FDR. For a phosphopeptide identification with discriminant score F_i , its q-value can be calculated:

$$q(F_i) = \min_{f \leq F_i} FDR(f) \quad (17)$$

Figure 2 illustrates the method used to calculate the q-values using the pdf of F . In this figure, the red and blue curves represent the pdfs of F of the incorrect and correct identifications, respectively. For any $f \leq F_i$, $FDR(f)$ is calculated:

$$FDR(f) = \frac{A_r}{A_b} \quad (18)$$

where A_r is the area (shaded in red) to the right of f and enclosed by the corrected pdf curve of the incorrect identifications. A_b is the area (shaded in blue) to the right of f and enclosed by the corrected pdf curve of the correct identifications. $A_r = E(V)$ and $A_b = E(R)$. The corrected pdf curves of the incorrect and correct identifications are represented by $\pi_0 p_0$ and $\pi_1 p_1$, respectively. For a comprehensive theoretical description of the *FDR* and q-value, see Storey and Benjamini.²²⁻²⁵

For each identification, the p-value and the posterior probability $p(+ | F)$ (i.e., the probability that the identification is correct given its discriminant score) can also be computed. The posterior probability can be calculated using Bayesian theory.⁸

$$p(+|F) = \frac{p(F|+) p(+)}{p(F|+) p(+) + p(F|-) p(-)} \quad (19)$$

where $p(F | +)$ and $p(F | -)$ are the pdfs of F for the correctly and incorrectly identified phosphopeptides, respectively, and $p(-)$ and $p(+)$ are the *a priori* probabilities that an incorrect and correct identification occurred, respectively.

Calculation of the composite *FDR*

The q-values that were calculated above were calculated from the precursor ion charge state-dependent discriminant score pdfs. Ultimately, the phosphopeptide identifications are filtered using a desired *FDR* for each charge state, and then a composite *FDR* is computed:

$$FDR = \frac{\sum_{i=1}^C N_i * FDR_i}{\sum_{i=1}^C N_i} \quad (20)$$

where the set of charge $i+$ identifications are filtered using FDR_i , N_i is the total number of charge state $i+$ filter-passing identifications, and $N_i * FDR_i$ is the expected number of charge state $i+$ filter-passing false-positive identifications.

Results

Application of this data analysis approach to a phosphoproteome dataset from the study of irradiated human skin fibroblasts resulted in a total of 11,004 top hit peptide identifications by SEQUEST. Peptide filtering removed un-modified peptides and peptides with charge states $\geq 4+$. High charge state peptides were removed because the number of such species was relatively small, and it was not possible to accurately estimate the distribution of the SEQUEST scores. For this particular dataset, only the charge state 2+ and 3+ phosphopeptides were considered for further analysis since there were no 1+ phosphopeptides (they were excluded from fragmentation during MS analysis). After the filtering, the total numbers of charge state 2+ and 3+ peptides were 3,159 and 3,215, respectively.

For each of the resultant 2+ and 3+ top hit phosphopeptides, the most likely incorporated phosphate arrangement was obtained using the peptide score, and the $\Delta Cn'$ for each true top hit was then calculated. Figure 3 is a histogram of the original $XCorr$ ranks of the true top (blue) and true second hits (red). This figure shows that the number of identifications that required the calculation of $\Delta Cn'$ was not negligible.

A scatter plot of $\overline{XCorr} - \overline{Sp} - \overline{\Delta Cn'}$ projected onto three 2-dimensional planes was produced (Figure 4). In each of these graphs, the lower-left and the upper-right clusters represent the incorrect and correct phosphopeptide identifications, respectively. A histogram of ΔM was produced (Figure 5) and used to determine the boundary of the central dense region. This information was used during the preliminary clustering of the identifications to calculate the initial values for EM. Note that it took ~70 iterations for the EM to converge (Figure 6).

With the joint pdfs estimated for the correct and incorrect identifications, a maximum likelihood method was used to determine the cluster membership. Figure 7 shows the probabilities p_- and p_+ that each identification belongs to cluster 0 and cluster 1, respectively. Any point that was above the 45° line was assigned to cluster 1, and any point below it was assigned to cluster 0.

The discriminant score weights were obtained using LDA for charge states 2+ and 3+, and the F of each identification was:

$$F = 0.6634 * \overline{XCorr} + 0.4092 * \overline{Sp} + 0.6265 * \overline{\Delta Cn'} \quad (2+)$$

$$F = 0.3671 * \overline{XCorr} + 0.6944 * \overline{Sp} + 0.6189 * \overline{\Delta Cn'} \quad (3+)$$

The histograms of F are shown in Figures 8A and B for charge states 2+ and 3+, respectively. EM was used to estimate the pdfs of F , and these are plotted as the continuous curves in Figures 8A and B. The p-value, q-value, and $p(+|F)$ of each peptide identification were computed and their relationship with F is plotted in Figures 8C and D. The false negative rate (FNR) vs. F is also plotted (in Figures 8C and D). For each discriminant score, the estimated FNR is the area to the left of the discriminant score and enclosed by the pdf curve of the correct identifications. Note that the p-value and q-value shared the same trend and $p(+|F)$ and FNR also share the same trend.

The phosphopeptide identifications were filtered by a range of q-values to produce a range of FDR values. The expected number of false positive and true positive identifications corresponding to each FDR was computed:

$$N_{FP} = N_{\text{claimed_correct}} * FDR \quad (23)$$

$$N_{TP} = N_{\text{claimed_correct}} * (1 - FDR) \quad (24)$$

The relationship between N_{TP} and N_{FP} , i.e., the Receiver Operating Characteristic (ROC), was plotted for both charge states (Figures 8 E and F). Both ROCs indicated that the incorrect and correct identifications were distinguished with high certainty.

To summarize, the data analysis procedure required $XCorr$, Sp , ΔCn , ΔM , and a peptide sequence as input, and it output the p-value, q-value, and $p(+|F)$. The data flow from each data analysis step to the next one was seamless. In particular, the EM algorithm took $\overline{XCorr}$, $\overline{Sp}$, $\overline{\Delta Cn'}$, and ΔM as input and output the mixture model (i.e. the estimated parameters $\pi_0, \pi_1, \mu_0, \mu_1, \Sigma_0$ and Σ_1), and the LDA algorithm took the $\overline{XCorr}$, $\overline{Sp}$, $\overline{\Delta Cn'}$, and the cluster membership (obtained via the maximum likelihood analysis) of each identification as input, and output the optimal weights $U = [u_1, u_2, u_3]^T$.

Validation of the *FDR* estimation

Validation of the *FDR* estimation algorithm was carried out on two different datasets. One of the dataset was the one used in this paper and the other was a published dataset that had been manually validated.²⁶ For the former dataset, the set of phosphopeptide identifications with a certain *FDR* (either 0.0001, 0.001, or 0.01) was randomly sampled, and the corresponding experimental MS/MS spectra were manually annotated. The true *FDR* was then determined based on the manual annotation, adjusted by taking into account the artifacts in the data processing of peptide search engines, and ultimately compared with the estimated *FDR*. The difference between them was negligibly small. For the latter dataset, the *FDR* estimation algorithm was applied and the list of peptide identifications that passed the q-value cutoff were compared with the list of confident peptide identifications provided in the published paper. It was found that almost all of the q-value filter-passing peptides were reported as confident and *vice versa*. The results from both of the validation studies demonstrated the value of the *FDR* estimation algorithm presented in this paper. Details pertaining to the validation and the corresponding data are provided as supplementary material.

Discussion

Justification of the Gaussian Mixture Model

In the estimation of the joint distribution of $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$ using EM, it was assumed that the distribution is jointly normal. In practice, a small deviation of the distribution from the GMM should not exert a large influence on the estimated *FDR*. This effect is because only the estimated pdf of $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$ are used to obtain the cluster membership of each phosphopeptide identification, and because larger values of $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$ will always result in a larger discriminant score.

The pdf of *F*, which was modeled using a GMM, affects the estimated *FDR* to a relatively larger degree. The *F* scores calculated by PeptideProphet⁸ for incorrect and correct identifications are assumed to follow a gamma and a normal distribution, respectively. A GMM was used here because the *F* scores of the incorrect identifications appeared to be normal based on a visual inspection.

Justification of the multiple hypothesis testing

Estimating the *FDR* is a multiple hypothesis testing problem; each phosphopeptide identification represents a single hypothesis testing problem. Ideally, each p-value would be estimated using the null hypothesis distribution of *F* specific to statistical samples of identifications of each phosphopeptide. Similarly, each q-value would be estimated using the null and alternative hypothesis distributions of *F* specific to statistical samples of identifications of each phosphopeptide. Because of the nature of shot-gun proteomics, statistical samples of each phosphopeptide identification is unobtainable. Therefore, as with PeptideProphet,⁸ the null distributions of *F* were assumed to be identical for all of the peptide identifications.

Uncertainty of peptide identifications, determination of the most likely phosphorylation site(s), and *FDR* estimation

There are several levels of uncertainty in identifying phosphopeptides by the algorithm presented in this paper. First, in the *FDR* estimation described above, it was assumed that higher values of *XCORR*, *Sp*, and $\Delta Cn'$ result in higher-confidence phosphopeptide identifications. While this assumption is very reasonable from a statistical point of view, counter examples have been found by manual validation.²⁷ Therefore, a peptide identification that passes a very low q-value cutoff might actually be a wrong identification due to an incorrect high-scoring

match between an experimental and a theoretical spectrum, which ultimately could be reflected as a deviation between the estimated and true *FDR*.

Second, the uncertainty of the *FDR* estimation can come from ΔCn . ΔCn represents the normalized difference of *XCorr* between the top and second hit and does not correlate directly with the goodness-of-fit between the experimental and theoretical spectra. Furthermore, ΔCn can be affected by the size of the protein database that the search engine searches against. The larger the size of the database, the smaller ΔCn tends to get.

Third, the most likely arrangement of the phosphate group(s) is determined by comparing the peptide scores of different arrangements. Even the most likely arrangement might not have a high-confidence peptide score.¹⁶ Therefore, a confident phosphopeptide identification could still have an ambiguous arrangement of phosphate group(s). When the number of possible arrangements of the phosphate group(s) is large, it is more difficult to distinguish the most likely arrangement from the rest due to closeness of the peptide scores between multiple different arrangements. For this reason, the number of phosphate groups per peptide was limited to 3, by filtering the data before the *FDR* analysis.

As a result of these uncertainties, the *FDR* estimation algorithm presented in this paper performs a statistical analysis of phosphopeptide identifications. Individual phosphopeptide identifications that are confident based on their q-values can still be ambiguous in terms of their arrangement of phosphate group(s). If the absolute confidence of the arrangement of phosphate group(s) is needed, the AScore¹⁶ that quantifies the confidence of each phosphorylation site assignment can be calculated.

Conclusion

The data analysis pipeline presented herein was designed to estimate the *FDR* of phosphopeptide identifications, and its utility was demonstrated in the context of a proteomics study of irradiated human skin fibroblasts. The pipeline obtains weights that optimally define a discriminant score and estimates the q-values and *FDR* by performing a rearrangement of the incorporated phosphates to identify true top hits, data training using EM to obtain the joint probability density function of the peptide scores, and performing a linear discriminant analysis. This pipeline could be extended to analyze the search results of unmodified peptides, as well as to search peptide scores from other peptide search engines (e.g., X! Tandem) and multiple search engines.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

Portions of this research were supported by the NIH National Center for Research Resources (RR018522; RDS) and the U.S. Department of Energy (DOE) Office of Biological and Environmental Research Low Dose Radiation Research Program. Experiments and data analyses were performed in the Environmental Molecular Sciences Laboratory, a DOE national scientific user facility located at the Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the DOE by Battelle under Contract DE-AC05-76RLO 1830.

References

1. Hunter T. Signaling--2000 and beyond. *Cell* 2000;100(1):113–27. [PubMed: 10647936]
2. Reed SI. G1/S regulatory mechanisms from yeast to man. *Prog Cell Cycle Res* 1996;2:15–27. [PubMed: 9552379]

3. Ciechanover A, Orian A, Schwartz AL. Ubiquitin-mediated proteolysis: biological regulation via destruction. *Bioessays* 2000;22(5):442–51. [PubMed: 10797484]
4. Zimmer JS, Monroe ME, Qian WJ, Smith RD. Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom Rev* 2006;25(3):450–82. [PubMed: 16429408]
5. Kall L, Storey JD, Maccoss MJ, Noble WS. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *J Proteome Res*. 2007
6. Kall L, Storey JD, Maccoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res* 2008;7(1):40–4. [PubMed: 18052118]
7. Choi H, Nesvizhskii AI. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J Proteome Res* 2008;7(1):47–50. [PubMed: 18067251]
8. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002;74(20):5383–92. [PubMed: 12403597]
9. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007;4(3):207–14. [PubMed: 17327847]
10. Eng JK, McCormack AL, Yates JR. 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.
11. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;20(9):1466–7. [PubMed: 14976030]
12. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* 2003;2(1):43–50. [PubMed: 12643542]
13. Yang F, Stenoien DL, Strittmatter EF, Wang J, Ding L, Lipton MS, Monroe ME, Nicora CD, Gristenko MA, Tang K, Fang R, Adkins JN, Camp DG 2nd, Chen DJ, Smith RD. Phosphoproteome profiling of human skin fibroblast cells in response to low- and high-dose irradiation. *J Proteome Res* 2006;5(5):1252–60. [PubMed: 16674116]
14. Wang Y, Ding SJ, Wang W, Jacobs JM, Qian WJ, Moore RJ, Yang F, Camp DG 2nd, Smith RD, Klemke RL. Profiling signaling polarity in chemotactic cells. *Proc Natl Acad Sci U S A* 2007;104(20):8328–33. [PubMed: 17494752]
15. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics* 2004;4(7):1985–8. [PubMed: 15221759]
16. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* 2006;24(10):1285–92. [PubMed: 16964243]
17. Lopez-Ferrer D, Martinez-Bartolome S, Villar M, Campillos M, Martin-Maroto F, Vazquez J. Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST. *Anal Chem* 2004;76(23):6853–60. [PubMed: 15571333]
18. Dempster AP, Laird NM, Rubin DB. Maximum-likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc* 1977;39Ser. B
19. Redner R, Walker H. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* 1984;26(2)
20. Xu L, Jordan MI. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural computation* 1996;8:129–51.
21. Fukunaga, K. Introduction to statistical pattern recognition. Academic Press; San Diego, California: 1990.
22. Storey JD. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* 2003;31:2013–35.
23. Storey JD. A direct approach to false discovery rate. *J. R. Statist. Soc. B* 2002;64:479–98.
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)* 1995;57(1): 289–300.

25. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 2003;100(16):9440–5. [PubMed: 12883005]
26. Yang F, Camp DG 2nd, Gritsenko MA, Luo Q, Kelly RT, Clauss TR, Brinkley WR, Smith RD, Stenoién DL. Identification of a novel mitotic phosphorylation motif associated with protein localization to the mitotic apparatus. *J Cell Sci* 2007;120(Pt 22):4060–70. [PubMed: 17971412]
27. Chen Y, Kwon SW, Kim SC, Zhao Y. Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. *J Proteome Res* 2005;4(3):998–1005. [PubMed: 15952748]

Figure 1. Flow chart of the data analysis pipeline used to estimate the *FDR* of phosphopeptide identifications.

Figure 2.

Illustration of the q-value calculation. The red and blue curves denote the pdfs of F that correspond to the incorrect and correct phosphopeptide identifications, respectively. For a given discriminant score f , $FDR(f)$ is the ratio of the red area divided by the blue area. The q-value is then calculated by Eqn. (17).

Figure 3.
Histogram of the *XCorr* rank of the true top hit (blue) and the true second hit (red).

Figure 4.

Scatter plots of the SEQUEST search results. The lower-left and upper-right clusters corresponded to the incorrect and correct phosphopeptide identifications, respectively. (A).

$\overline{\Delta Cn'}$ vs. $\overline{XCorr}$. (B). $\overline{\Delta Cn'}$ vs. $\overline{Sp}$. (C). $\overline{Sp}$ vs. $\overline{XCorr}$.

Figure 5. Histogram of ΔM . The maximum lower and minimum upper thresholds of the central dense region were 0 and 10 ppm, respectively.

Figure 6.

Estimation of the joint pdf of $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$ using EM. The x-axis denotes the indices of iterations and the y-axis denotes the values of the estimated parameters. The parameters that correspond to the incorrect and correct identifications are red and blue, respectively. (A). Convergence of π_0 and π_1 . (B). Convergence of $\mu_{0,\overline{XCorr}}$ and $\mu_{1,\overline{XCorr}}$. (C). Convergence of $\mu_{0,\overline{Sp}}$ and $\mu_{1,\overline{Sp}}$. (D). Convergence of $\mu_{0,\overline{\Delta Cn'}}$ and $\mu_{1,\overline{\Delta Cn'}}$. (E) Convergence of $\text{cov}(\overline{XCorr}, \overline{\Delta Cn'})$. (F). Convergence of $\text{cov}(\overline{Sp}, \overline{\Delta Cn'})$. (G). Convergence of $\text{cov}(\overline{XCorr}, \overline{Sp})$. The insets in (B), (C), and (D) are the estimated means of $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn'}$ on a zoomed-in scale.

Figure 7.

Determination of the cluster membership using the joint pdf of $\overline{XCorr}$, $\overline{Sp}$, and $\overline{\Delta Cn}$. The x- and y-axes denote the probability that each phosphopeptide identification belongs to the distribution of incorrect (p^-) and correct (p^+) identifications, respectively. The identifications in red (below the 45° line) were assigned to cluster 0, and those in blue (above the 45° line) were assigned to cluster 1.

Figure 8.

A,B: Histogram of F and the estimated pdf of F . The red and blue curves correspond to the incorrect and correct identifications, respectively. C,D: The p-value (red), q-value (green), and $p(+|F)$ (blue) for each identification. The FNR is shown in cyan. E,F: ROC curves.

Table 1Possible outcomes from thresholding N peptide identifications for significance

	declared non-significant	declared significant	total
null hypothesis true	O	V	Z
alternative hypothesis true	T	S	$N-Z$
total	$N-R$	R	N