Improving phosphopeptide identification in shotgun proteomics by supervised filtering of peptide-spectrum matches

Sujun Li,[1] Randy J. Arnold,[2] Haixu Tang,[1] Predrag Radivojac[1]
1) Department of Computer Science and Informatics; 2) Department of Chemistry
Indiana University, Bloomington, Indiana, U.S.A.

ABSTRACT
One of the important objectives in mass spectrometry-based proteomics is the identification of post-translationally modified sites in cellular and extracellular proteomes. Proteomics techniques have been particularly effective in studying protein phosphorylation, where tens of thousands of new sites have been recently discovered in all domains of life. Such massive discovery of new sites has been facilitated by progress in affinity enrichment techniques, high-throughput analytical platforms that couple liquid chromatography (LC) and tandem mass spectrometry (MS/MS), and also powerful computational tools that assign peptides to tandem mass spectra. In this work we focus on computational protocols for identifying phosphoproteins, phosphopeptides, and phosphosites. Although the current tools already provide solid results, most methods have not been tuned to exploit particular sequence and physicochemical properties of phosphopeptides or the peculiarities of their fragment spectra. Therefore, novel algorithms can be designed to increase the sensitivity of phosphosite identification. Here we describe a machine learning-based method that improves the identification of phosphopeptides in LC-MS/MS experiments. Our algorithm is applied as a post-processing step to a standard database search. It assigns a probability score to each peptide-spectrum match (PSM) corresponding to a phosphopeptide, based on the sequence and spectral features of the peptide and its assigned fragment spectra as well as the biological propensity of particular residues in the peptide to be phosphorylated. The algorithm is based on a simple but robust logistic regression model and is used together with a conventional search engine (here, MASCOT) to filter out the PSMs with the lowest probability of being correctly identified. Our protocol was tested on two large phosphoproteomics data sets on which it increased the number of identified phosphopeptides by 10-15% compared to the conventional scoring algorithms at the same false discovery rate threshold of 1%.

Categories and Subject Descriptors
I.5.0 [**Pattern Recognition**]: General; J.2 [**Life and Medical Sciences**]: Biology and genetics.

BCB '13, September 22 - 25, 2013, Washington, DC, USA

Keywords
Bioinformatics, proteomics, tandem mass spectrometry, post-translational modification, phosphorylation, machine learning.

1. INTRODUCTION
As one of the major and most common protein post-translational modifications (PTMs), phosphorylation has been studied extensively [1]. Numerous efforts have been made to understand the structural [2-6], evolutionary [7, 8], and functional aspects of phosphorylation [9, 10]. Several major repositories have been created to store and characterize phosphorylation sites across multiple organisms [11, 12], and a number of computational models dedicated to predicting phosphorylation sites have been developed [13]. Due to its importance in cellular signal transduction and functional regulation, links between dysregulation of phosphorylation and disease have been established both experimentally and computationally. Examples include a number of monogenic disorders [14, 15], but also complex diseases such as cancer [3, 16, 17], heart disease [18], and neurodegenerative disorders [19].

Tandem mass spectrometry (MS/MS) has been the most popular high-throughput experimental approach for identifying phosphopeptides and advancing the understanding of the function and dynamics of protein phosphorylation [20-22]. In these studies, MS/MS spectrum data are often analyzed by database search engines such as MASCOT [23] and SEQUEST [24]. Although these tools yield reasonably good performance in phosphopeptide identification, they were originally designed for the identification of non-modified peptides and thus do not necessarily exploit specific sequence characteristics of phosphopeptides [25] and their fragmentation patterns in tandem mass spectrometers [26]. Thus, advanced computational methods are highly desirable to improve phosphopeptide identification from the massive MS/MS data sets.

To date, significant efforts have been made to improve the identification of phosphopeptides and phosphosites. For instance, Lu et al. [27] developed a machine-learning method to filter MS/MS spectra corresponding to phosphopeptides prior to database searching. Novel scoring methods optimized specifically for phosphopeptide-spectrum matching were also developed. Payne et al. [28], for example, used a Bayesian network to incorporate fragmentation rules learned from the MS/MS spectra of phosphopeptides. Other methods also employed machine-learning algorithms to increase the number of identified phosphopeptides in the post-processing of phosphopeptide-spectrum matches (pPSMs) reported by search engines [29, 30]. In addition to phosphopeptide identification, various groups have focused on accurate identification of phosphory-

lation sites. While this process is trivial for the peptides containing one phosphorylatable residue, it is far more difficult in a general case. Computational methods that differentiate between several potential phosphorylation sites based on the identified MS/MS spectra have also been developed [20, 31-33]. Among those, the A-score [32] algorithm, a simple yet powerful method for the probabilistic assignment of phosphorylation sites has been commonly used.

One drawback of all existing phosphopeptide identification algorithms is that they do not consider an important and prominent feature of phosphorylation. They treat all phosphorylatable residues (i.e., Ser, Thr, or Tyr) equally, yet it has been shown that consensus peptide sequence patterns or patterns of physicochemical properties around phosphorylation sites can be observed [25, 34]. Although these patterns are highly diverse when regulated by different kinases [35] or in different species, successful bioinformatics approaches have been developed to predict the propensity of a residue to be phosphorylated [36, 37]. These predictors have been applied to a number of interesting biological problems [38-40]; however, to our knowledge, their application to mass spectrometry-based proteomics data analysis has not been previously explored.

In this study we present a new algorithm that incorporates sequence, spectral, and biological signatures of pPSM and a phosphorylation site into the post-processing of the results from database search engines. Our method, named PhosART, employs a supervised learning algorithm (logistic regression) to estimate the probability that a pPSM is correct in a MS/MS experiment. By testing PhosART on two large phosphoproteomics data sets and comparing its performance to that of standard database search, we show that it increases the number of identified phosphopeptides at the same false discovery rate (FDR) thresholds.

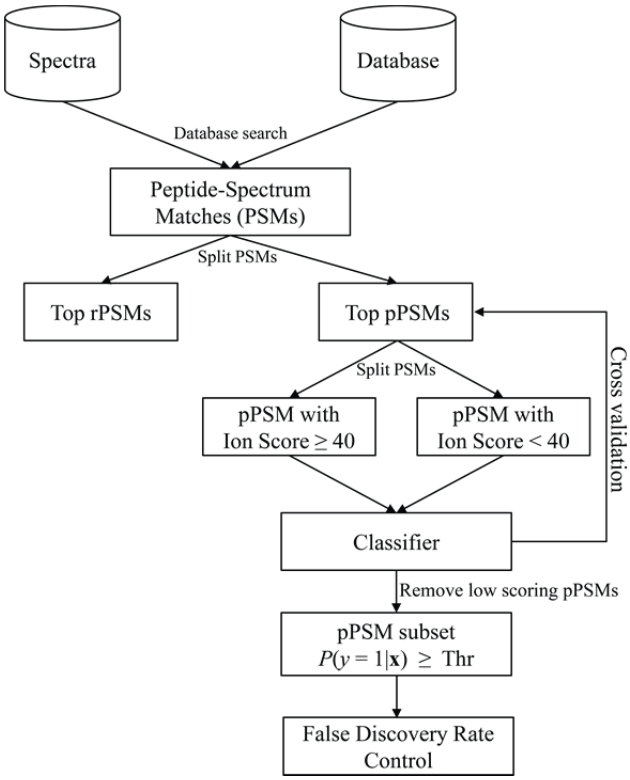

Figure 1. Illustration of the workflow for phosphopeptide identification. rPSMs represent regular (non-phosphorylation) PSMs, while pPSMs represent phosphopeptide-spectrum matches.

2. METHODS

2.1 Data sets

In this work we used two data sets in which affinity enrichment was used for improving phosphopeptide identification.

Data set I contains proteomics data obtained from Iliuk et al. [41], who devised a method called polymer-based metal-ion affinity capture (PolyMAC) using water-soluble dendrimers covered with phosphopeptide-binding titanium molecules for phosphopeptide analysis. The human proteomic samples were enriched for phosphopeptides by their in-house PolyMAC reagents and then analyzed with MS/MS by coupling high-performance liquid chromatography (HPLC) with a high-resolution hybrid linear ion trap Orbitrap mass spectrometer (LTQ-Orbitrap XL, Thermo Fisher Scientific). The mass spectrometer was operated in the data-dependent mode in which a full MS scan was followed by four MS/MS scans of the most abundant ions. Ions with the charge state of +1 were excluded.

Data set II was acquired from the analysis of human samples by Kim et al. [42]. Here, a phosphopeptide mixture was extracted from pancreatic cancer cell line followed by a trypsin digestion. The digested sample was subsequently fractionated on the Strong Cation Exchange (SCX) column, and the phosphopeptides were enriched using TiO_2. The mass spectrometry experiments were performed on an LTQ-Orbitrap XL ETD, a hybrid instrument combining collision-induced dissociation (CID), electron-transfer dissociation (ETD), and pulsed-Q dissociation (PQD). The data set was downloaded from the Tranche network at `http://proteomecommons.org`. Detailed description of the experimental protocol can be obtained from the original study [42]. In this work we only utilized CID data.

Both data sets were searched against all human proteins from the UniProtKB database (release 2013_04) using MASCOT v.2.4. The search database contained a combined forward and reverse human proteome with the following search criteria: (*i*) tryptic enzyme specificity with at most one missed cleavage; (*ii*) carbamidomethylation at cysteine residues as a fixed modification; (*iii*) phosphorylation at serine, threonine, and tyrosine, and oxidation at methionine residues as variable modifications; (*iv*) 1.0 Da for precursor ion tolerance and 0.5 Da for fragment ion tolerance. Peptide mass calculation was performed using monoisotopic values.

2.2 Data processing workflow

Peptide identification was performed in several steps, as shown in Figure 1. All top-scoring peptide-spectrum matches (PSMs) from the database search were grouped into unmodified PSMs and phosphorylated PSMs. The PSMs corresponding to the unmodified peptides were subsequently removed, and the remaining PSMs were considered for the further analysis of phosphopeptide-spectrum matches (pPSMs).

The next step consisted of a procedure that filters out a subset of top-scoring pPSMs that were least likely to be correct identifi-

cations. This step was carried out via a 10-fold cross-validation, prior to the FDR-based identification of confident PSMs (i.e. those with FDR below a particular stringent threshold). First, the entire set of pPSMs was partitioned into 10 equal-sized non-overlapping sets. In each of the 10 iterations of cross-validation, a classification model was then trained on 9/10-ths of the data and applied to the remaining 1/10-th. Once a score for each of the pPSMs (in the entire data set) was provided, those with the lowest scores were simply removed from the list.

To train a classifier, all pPSMs with a MASCOT ion score ≥40 were used in the positive set, whereas the remaining pPSMs from the forward (target) database were considered as the negative set. The resulting training data was highly imbalanced in both data sets; there were 2225 positive pPSMs and 14283 negative pPSMs in data set I along with 1953 positive and 13161 negative pPSMs in data set II. In order to balance positive and negative sets and still exploit all negatives, we trained a logistic regression model found through maximization of the weighted likelihood function. In particular, given a data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i$ is a column vector of features (describing peptides) and $y_i \in \{0,1\}$ is the class variable, the weighted likelihood function can be expressed as

$$l(\mathbf{w}) = \prod_{i=1}^{n} p_i^{c_i y_i} (1 - p_i)^{c_i (1 - y_i)}$$

where $\mathbf{w}$ is a set of weights of the logistic regression model, $p_i = P(y_i = 1|\mathbf{x}_i) = 1/\left(1 + e^{-\mathbf{w}^T \mathbf{x}_i}\right)$ is the output of the logistic regression classifier for data point $\mathbf{x}_i$, and $0 \leq c_i \leq 1$ is a cost (weight) assigned to data point $\mathbf{x}_i$.

The weight update rule for this classifier can be derived using the standard Netwton-Raphson technique in maximizing the log-likelihood function $ll(\mathbf{w}) = \log l(\mathbf{w})$ as

$$\mathbf{w} \leftarrow \mathbf{w} + H_{ll(\mathbf{w})}^{-1} \cdot \nabla ll(\mathbf{w})$$

where $\nabla ll(\mathbf{w})$ is the gradient and $H_{ll(\mathbf{w})}$ is the Hessian matrix of $ll(\mathbf{w})$. It is relatively straightforward to derive that

$$\nabla ll(\mathbf{w}) = \mathbf{X}^T \mathbf{C}(\mathbf{y} - \mathbf{p})$$

and

$$H_{ll(\mathbf{w})} = -\mathbf{X}^T \mathbf{C}(\mathbf{I} - \mathbf{P})\mathbf{X}$$

where $\mathbf{X}$ is the data matrix (augmented by a column of ones to prevent the separating hyperplane from passing through the origin of the coordinate system), $\mathbf{y}$ and $\mathbf{p}$ are the column vectors of class variables and their predictions on the training set, $\mathbf{c}$ is the column vector of costs, $\mathbf{P} = \text{diag}\{\mathbf{p}\}$, $\mathbf{C} = \text{diag}\{\mathbf{c}\}$, and $\mathbf{I}$ is the identity matrix. Because all values in $\mathbf{p}$ and $\mathbf{c}$ are non-negative and less than or equal to one, the Hessian matrix is negative semi-definite. This means that the set of weights $\mathbf{w}$ found through this procedure globally maximizes the weighted likelihood function. The initial set of weights can be found using a weighted ordinary least-squares regression [43] as

$$\mathbf{w} = (\mathbf{XCX})^{-1} \mathbf{X}^T \mathbf{Cy}.$$

To ensure that the total costs for positive and negative data sets were balanced, the costs for all positive data points were set to one and the costs for negative points were set to the ratio of positive and negative data points in the training set. Note that before learning the weights of the logistic regression model, we normalized the data using the *z*-score normalization and subsequently performed principal component analysis, with 99% of the variance retained, to remove (nearly) co-linear features and ensure invertibility of the Hessian. Both operations were performed on the training set only and then applied on the test set.

It is important to mention that the rationale for choosing a relatively high MASCOT score threshold to construct positive examples was to ensure that the positives almost entirely consisted of correct pPSMs, while the set of negatives was allowed to be noisy. A linear classifier such as logistic regression is well suited to be trained from such data.

Once low-scoring pPSMs are eliminated from the database, the final step of the data processing was a standard target-decoy procedure that was applied to output a set of confidently identified spectra. The prediction threshold that provided the most confident phosphopeptide identifications at a particular FDR was selected. Target-decoy-based FDR estimation was based on MASCOT scores only.

2.3 Feature representation

The features used to represent each PSM can be categorized into three groups: (*i*) sequence-based features, (*ii*) spectral features, and (*iii*) biochemical features. The sequence-based features generated for each peptide sequence contained amino acid compositions, peptide length, peptide mass, and the counts of putative phosphorylation sites (serine, threonine, and tyrosine residues) in the peptide. In total, 25 sequence-based features were generated for each peptide sequence. The spectral features included: the number of fragment ion pairs with an m/z difference of 98, 80, and 49 (doubly charged); the intensity ratio between the precursor ion with 98 Da neutral loss and the base peak; intensity ratio between the precursor ion with a neutral loss of 18 Da and the base peak; the percentage of peaks with intensity above 1% of the total intensity; and the intensity difference between the highest and second highest peak. Finally, the biochemical features used in the machine-learning model included the predicted phosphorylation propensity by using the DisPhos phosphorylation site predictor [3]. DisPhos is a multi-species, kinase-independent predictor of phosphorylation sites that exploits local amino acid sequence profiles as well as predicted structural properties of the residue neighborhood (e.g. intrinsically disordered regions). These features included the mean and maximum DisPhos scores in the peptide among multiple S/T/Y sites. It is worth noting that the DisPhos features were generated when the entire protein sequence was used for the prediction of phosphorylation sites (residues outside of the tryptic peptide sequence were also exploited). Thus, the same peptide can contain different phosphorylation scores if present in different proteins or species, depending upon its flanking regions.

3. EXPERIMENTS AND RESULTS

3.2 Accuracy of classification models

To characterize the accuracy and thoroughly evaluate classification models, we estimated the area under the Receiver Operating Characteristic (ROC) curve for the logistic regression

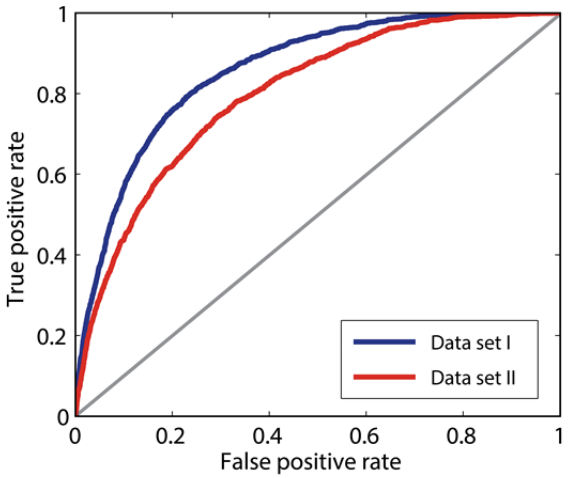

Figure 2. ROC curve for the trained classifiers. The areas under the ROC curve were estimated to be 0.86 and 0.80 for data set I and data set II, respectively. The straight gray line shows the accuracy of a random classifier.

classifiers on each of the data sets. The ROC curve displays true positive rate of classification as a function of false positive rate and was calculated over a set of decision thresholds. For data set I, we estimated the area under the ROC curve (AUC) to be 0.86. Similarly, for data set II we estimated an AUC of 0.80. Both ROC curves are shown in Figure 2. It is important to mention that while the set of positive data points can be considered to be of high quality, there may be a significant fraction of negative data points that were incorrectly assigned class in our data set construction. Therefore, the true accuracy of the classifier is likely to be higher than estimated.

3.1 Target-decoy search of the filtered PSMs

In order to report the result of phosphopeptide identification at a given FDR, PhosART combines the predicted probability of correct identification and MASCOT ion scores. The integration method simply removes the pPSMs with low scores by the machine-learning classifier. Based on the retained subset of pPSMs, the target-decoy approach was subsequently applied to the original MASCOT scores to report the actual identification result at a given false discovery rate.

Figure 3A shows the number of retained pPSMs at different probability thresholds for both data sets. Similarly, Figure 3B shows the number of phosphopeptide identifications at given FDR thresholds (1% and 5%). To obtain stable estimates, the numbers in each figure were averaged over 10 different 10-fold cross-validation runs, each with a different partitioning. Interestingly, the number of confidently identified pPSMs was not reduced with the decreasing pool of available pPSMs. Specifically, the number of phosphopeptide identifications exhibits a unique distribution pattern. For data set I, at the prediction threshold of $\tau = 0.2$ (i.e. when pPSMs with scores below 0.2 were permanently removed from the pPSM pool), approximately 30% of pPSMs were removed from the pool, yet the number of identified peptides increased by 14.6% (4745 vs. 4142). Similarly, the removal of pPSMs with scores below $\tau = 0.3$ increased the number of identified phosphopeptides in data set II by 10.3% (3815 vs. 3460). The threshold providing the most identifications was relatively consistent across the two data sets.

These results are also summarized in Table 1 and analyzed for statistical significance. Given a particular range of scores (defined by the threshold value τ), the number of identified peptides can be considered to follow a Gaussian distribution with mean μ and standard deviation σ (e.g. for $\tau = 0.1$ and data set I, $\mu_I(0.1) = 4512.1$ and $\sigma_I(0.1) = 66.7$). Therefore, the probability that the number of identified peptides is less than or equal to 4142 (for data set I) or 3460 (for data set II) corresponds to the value of the cumulative distribution function of the Gaussian distribution with parameters μ and σ at point 4142 (data set I) or 3460 (data set II). This probability corresponds to the P-value that the proposed algorithm provides an increase in the number of identified phosphopeptides.

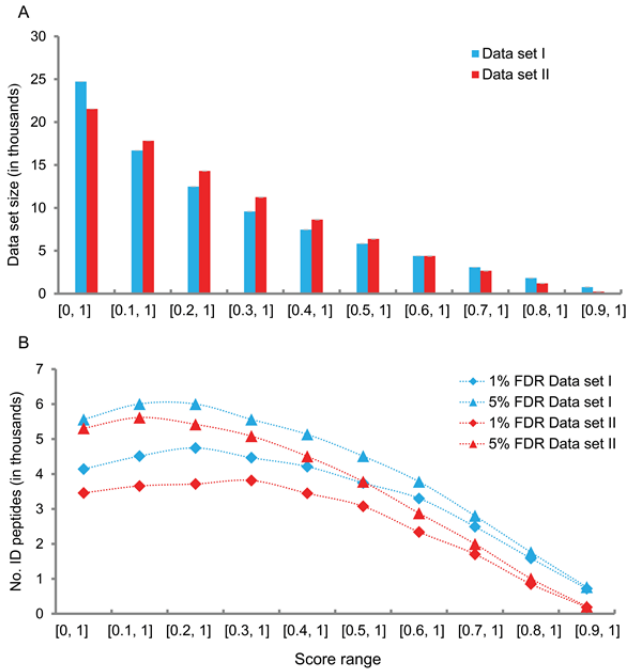

Figure 3. A. The number of retained pPSMs for different probability thresholds. The x-axis shows the interval of scores used to retain pPSMs. B. The number of identified phosphopeptides as a function of the score threshold used to retain pPSMs, at 1% and 5% FDR levels.

3.3 Impact of different types of features

To evaluate the impact of each category of features, we constructed five predictors, each based on a different feature set. Namely, we separately evaluated sequence features, spectral features, and the features derived from the predicted propensity of phosphorylation sites. Strictly for comparison purposes, we also implemented a model that assigned probability scores from a uniformly random distribution. Top-scoring pPSMs were selected at the FDR cutoffs of 1% and 5%, using the same protocol described above.

The results corresponding to each of the above-mentioned predictors are shown in Figure 4. The baseline model identified 4142 (1% FDR) and 5561 (5% FDR) pPSMs in data set I. In

data set II, it identified 3460 (1% FDR) and 5308 (5% FDR) pPSMs, respectively. The random model identified 3715 (1% FDR) and 4998 (5% FDR) pPSMs in data set I as well as 3132 (1% FDR) and 4811 (5% FDR) pPSMs in data set II. As expected, the reduction in performance of a random model was approximately 10%, which matches the percentage of peptides removed when they are assigned prediction scores from a uniformly random distribution.

Table 1. The number of identified phosphopeptides (and standard error) averaged over ten different 10-fold cross-validation runs, each with a different partitioning. The threshold column (Thr) indicates the interval of PhosART scores for the retention of pPSMs in the data set (i.e. the pPSMs with scores below the lower limit were removed). The * marks in the table correspond to the situations with significant P-values, in each case below 10^{-13}). The remaining entries resulted in P-values between 0.02 and 1.

	Data set I	Data set II
Thr	No. identified peptides	No. identified peptides
[0,1]	4142.0	3460.0
[0.1,1]	4512.1 ± 21.1 *	3658.1 ± 1.7 *
[0.2,1]	**4745.1 ± 9.5** *	3714.1 ± 2.8 *
[0.3,1]	4466.7 ± 9.9 *	**3814.6 ± 15.4** *
[0.4,1]	4211.2 ± 10.9	3448.6 ± 10.3
[0.5,1]	3743.4 ± 11.5	3074.8 ± 8.2
[0.6,1]	3301.4 ± 16.5	2339.3 ± 9.7
[0.7,1]	2493.5 ± 18.7	1704.2 ± 10.4
[0.8,1]	1587.4 ± 6.5	851.7 ± 4.7
[0.9,1]	720.4 ± 9.9	188.7 ± 1.5

Figure 4 also shows that the PhosART predictor significantly increases the number of identified pPSMs. As specified, when using multiple categories of features, PhosART identified more pPSMs than by using only individual categories of features. However, the contribution of each category of features varies. When only spectral features or sequence features were used to train the predictor, fewer peptides were identified at the same FDR level than by both sequence and spectral features or all three categories of features together (Figure 4).

While the features derived from the predicted propensity of phosphorylation have not improved performance when applied in isolation, they had a positive effect when combined with sequence and spectral features. For example, in data set I, at 1% FDR level, 4142 phosphopeptides were identified by the baseline model, 4129 were identified when using sequence features, 4185 using spectral features, 4599 using sequence and spectral features, and 4745 by using all features. In data set II, at 1% FDR level, 3460 phosphopeptides were identified by the baseline model, 3629 when using sequence features, 3506 when using spectral features, 3796 using sequence and spectral features, and 3815 by using all features.

We also compared the phosphopeptides identified by the baseline and the PhosART approach and found that a large fraction of identified phosphopeptides between these two methods is common, ranging between 75% and 82%. The peptides identified by PhosART covered 91% and 93% of those identified by the baseline model. At the level of unique peptides, in data set I 251 (21.4%) more unique peptides were identified, while 150 (10.7%) more unique peptides were identified in data set II.

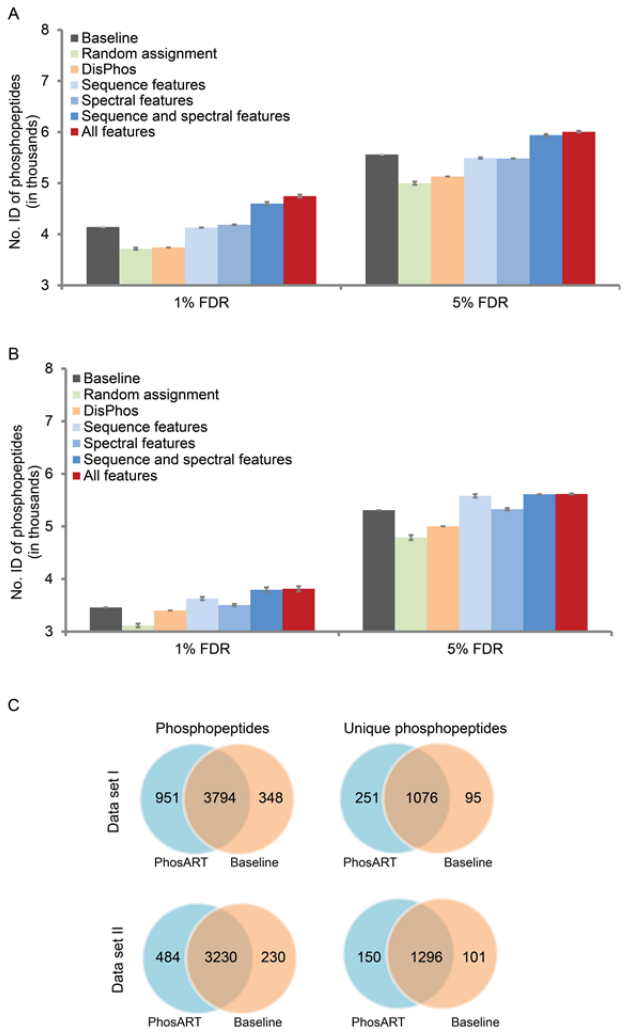

Figure 4. Phosphopeptide identification from different predictors based on individual features in A) data set I and B) data set II. The x-axis represents the probabilities assigned by different predictors, whereas the y-axis represents the number of identified phosphopeptides at 1% FDR or 5% FDR. C. The Venn diagram showing the number of identified phosphopeptides according to the baseline model and PhosART.

3.3 Impact of phosphorylation prediction

To further demonstrate the effect of the phosphorylation propensity predicted by DisPhos, we investigated the distribution of DisPhos scores amongst different subsets of pPSMs, as shown in Figure 5. Six subsets were constructed, including phos-all, phos-fdr1, phos-other, nophos-all, nophos-fdr1, and nophos-other. The phos-all data set included all pPSMs reported by MASCOT (Figure 1). The phos-fdr1 data set was the set of confidently identified pPSMs from phos-all, at the FDR cutoff of 1%. The phos-other data set contained the remaining pPSMs in phos-all excluding the ones in phos-fdr1. Similarly, the nophos-all set included all PSMs of the non-modified peptides. The nophos-fdr1 was a subset of nophos-all PSMs that included those identified at 1% FDR cutoff. The nophos-other subset contained the remaining PSMs in nophos-all. We note that only peptides with S/T/Y residues were used in the nophos data set. The mean and maximum DisPhos scores

over all phosphorylatable sites were calculated for each peptide in these six data sets.

Figure 5 shows the box plots of the maximum DisPhos scores in data set I (A) and data set II (B). Compared to the other subsets, phos-fdr1 set exhibits the highest average of maximum DisPhos scores. The mean and maximum of DisPhos scores follow the same trend. The significant difference between the confidently identified phosphopeptides and non-phosphopeptides ($P = 8.1 \cdot 10^{-49}$ and $P = 5.0 \cdot 10^{-29}$; t-test) in both data sets indicates that the sequence pattern around the phosphorylation sites could be utilized here to discriminate phosphopeptides with high propensity to be phosphorylated from other phosphopeptides and thus to reduce the probability that pPSMs are incorrect.

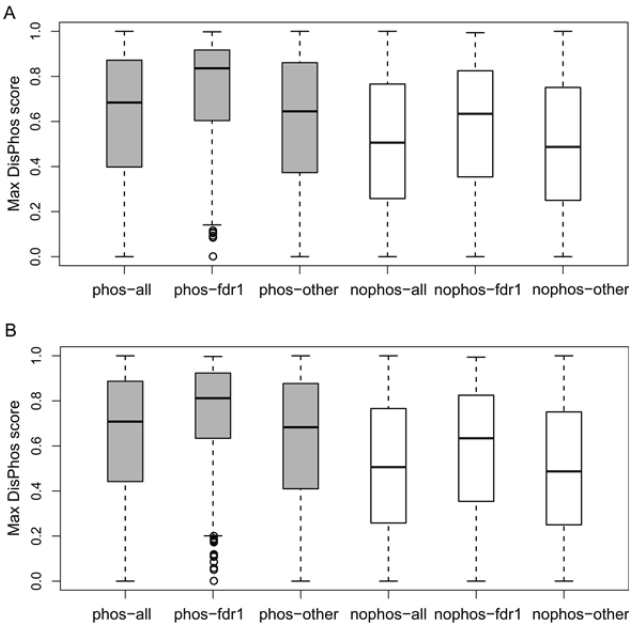

Figure 5. Box plot of DisPhos scores in different subsets of pPSMs: (A) for data set I and (B) for data set II. The x-axis shows the data subset. The y-axis represents the distribution of maximum scores over all phosphorylatable residues.

3.4 Quality of site localization

In PhosART we used logistic regression to assign a probability score to each pPSM. Through combining this probability score and MASCOT ion score, 10-15% more phosphopeptides were identified. To investigate the quality of novel sites identified by PhosART, we utilized the A-score algorithm to determine site localization, i.e. to identify those peptides/sites with A-scores greater than or equal to 19 [32] in the set of pPSMs that were identified by PhosART. Peptides with a single phosphorylatable site were ignored in this analysis.

As shown in Figure 6A, by using the baseline model (at 1% FDR), 66.2% of peptides in data set I and 58.3% in data set II had A-scores of at least 19. This percentage stayed roughly unchanged in the PhosART results where 64.3% of phosphopeptides in data set I and 57.0% in data set II had A-scores of 19 and above. In comparison, running the A-score algorithm on the randomly selected pPSMs from data set I returned only around 40% pPSMs with score 19 and higher. This implies that the low quality pPSMs may have a low chance of receiving high A-scores. We further investigated the fraction of predicted S/T/Y sites in the identified phosphopeptides as reported by two different models, baseline and PhosART. The results indicate that the change is minimal (Figure 6B) which suggests that PhosART model equally captured the sequence and spectral patterns for each different residue and did not alter the biologically innate S/T/Y distribution.

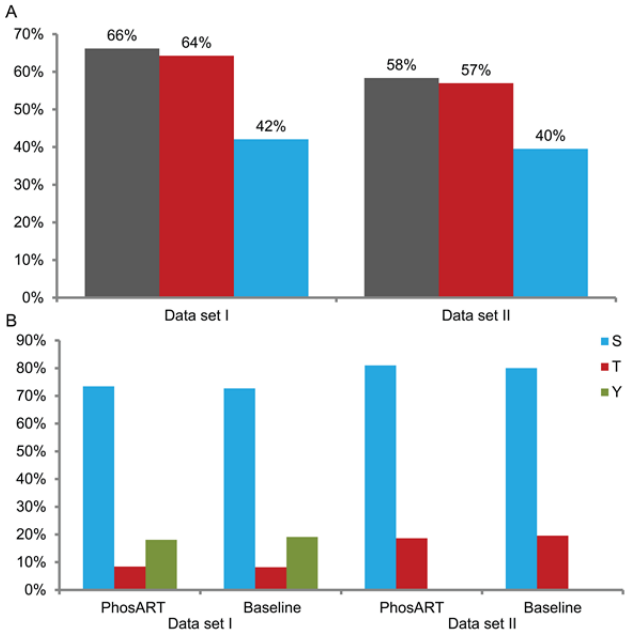

Figure 6. A. The comparison of the percentage of confidently localized phosphorylation sites (i.e., with A-score ≥19) in the selected pPSMs based on the PhosART probability scores (red), the baseline scores (gray) and randomly selected PSMs (blue) in data sets I and II. B. The percentage of S/T/Y residues in the baseline and PhosART identifications.

4. DISCUSSION

In this paper we describe a post-processing method developed to improve phosphopeptide identification in shotgun proteomics. The method employs diverse feature sets to construct a logistic regression predictor that assigns a probability score to each pPSM. The rationale for this method is that it simply exploits signatures of pPSMs that current database searching engines do not incorporate into the matching algorithm. By utilizing a range of sequence, spectral, and biological features, our method identified 10-15% more phosphopeptides at the same false discovery rate (FDR) as common search engines such as MASCOT.

PhosART conducts online training and prediction of pPSMs on a set of pPSMs that were identified by any peptide search engine, and thus does not rely on any prior training. In addition, the training data used by PhosART does not contain pPSMs from decoy peptides. As a result, the same target-decoy search can be used to estimate FDR in phosphopeptide identification. Thus, PhosART is ready to be used to improve phosphopeptide identification without a significant computational overhead. At this stage, PhosART exploits the fact that current search engines do not incorporate all relevant features for identification of phosphopeptides. As those search engines improve, the need for this post-processing technique is expected to diminish.

In training our model we decided to use simple linear classification. While, in principle, non-linear models such as neural networks or support vector machines could be employed in this step, linear classification with weighted likelihood has many desirable properties, such as applicability to small data sets or data sets with much higher class imbalance than observed in this work. The weighted likelihood function improved the area under the ROC curve of the logistic regression classifier by 1-2% in our experiments (data not shown). While this is a small improvement in its ranking ability, this method provides more meaningful scores from the classifier and naturally addresses the class imbalance problem. To ensure invertibility of the Hessian, we applied normalization and principal component analysis to the original data set. Although we have not tried it in this work, a better effect on classifier performance could be achieved using regularization in the optimization step.

In order to provide stable FDR estimates, we only utilized data sets where phosphopeptide affinity enrichment was applied. However, this method can also be used for any proteomic search in order to explore the existence or quantity of phosphopeptides. In one of such preliminary analyses we found that the predicted biological features (by DisPhos) provided more substantial improvements (data not shown). Therefore, we believe that there exist differences between enriched and non-enriched data sets and that all features used in this work are useful. In principle, there are no obstacles in exploiting the knowledge that a particular site has already been observed to be phosphorylated (instead of predicting the propensity for phosphorylation); however, we decided to only explore prediction of phosphorylation sites.

Finally, although this work only provides a proof of concept in several aspects of phosphopeptide identification, it is significant because it suggests that similar techniques can be applied to other post-translational modifications. Such studies will be more realistic as the databases of identified post-translationally modified sites become larger and their fragmentation spectrum patterns better understood.

5. ACKNOWLEDGMENTS

We thank Prof. W. Andy Tao of Purdue University for providing us with phosphorylation MS/MS data. This work was supported by the National Institutes of Health grant R01 GM103725-04 and National Science Foundation grant DBI-0644017.

6. REFERENCES

1. Walsh, C.T., *Posttranslational modification of proteins: expanding nature's inventory*. 2006, Roberts and Co. Publishers.
2. Gsponer, J., et al., Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science*, 2008. **322**(5906): 1365-8.
3. Iakoucheva, L.M., et al., The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*, 2004. **32**(3): 1037-49.
4. Johnson, L.N. and R.J. Lewis, Structural basis for control by phosphorylation. *Chem Rev*, 2001. **101**(8): 2209-42.
5. Xin, F. and P. Radivojac, Post-translational modifications induce significant yet not extreme changes to protein structure. *Bioinformatics*, 2012. **28**(22): 2905-13.
6. Nussinov, R., et al., Allosteric post-translational modification codes. *Trends Biochem Sci*, 2012. **37**(10): 447-55.
7. Landry, C.R., E.D. Levy, and S.W. Michnick, Weak functional constraints on phosphoproteomes. *Trends Genet*, 2009. **25**(5): 193-7.
8. Gray, V.E. and S. Kumar, Rampant purifying selection conserves positions with posttranslational modifications in human proteins. *Mol Biol Evol*, 2011. **28**(5): 1565-8.
9. Gnad, F., et al., High-accuracy identification and bioinformatic analysis of *in vivo* protein phosphorylation sites in yeast. *Proteomics*, 2009. **9**(20): 4642-52.
10. Gnad, F., et al., Evolutionary constraints of phosphorylation in eukaryotes, prokaryotes, and mitochondria. *Mol Cell Proteomics*, 2010. **9**(12): 2642-53.
11. Gnad, F., J. Gunawardena, and M. Mann, PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res*, 2011. **39**(Database issue): D253-60.
12. Bodenmiller, B., et al., PhosphoPep-a database of protein phosphorylation sites in model organisms. *Nat Biotechnol*, 2008. **26**(12): 1339-1340.
13. Eisenhaber, B. and F. Eisenhaber, Prediction of posttranslational modification of proteins from their amino acid sequence. *Methods Mol Biol*, 2010. **609**: 365-84.
14. Li, S., et al., Loss of post-translational modification sites in disease. *Pac Symp Biocomput*, 2010: 337-47.
15. Mort, M., et al., *In silico* functional profiling of human disease-associated and polymorphic amino acid substitutions. *Hum Mutat*, 2010. **31**(3): 335-46.
16. Radivojac, P., et al., Gain and loss of phosphorylation sites in human cancer. *Bioinformatics*, 2008. **24**(16): i241-7.
17. Li, B., et al., Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, 2009. **25**(21): 2744-2750.
18. Hata, J.A. and W.J. Koch, Phosphorylation of G protein-coupled receptors: GPCR kinases in heart disease. *Mol Interv*, 2003. **3**(5): 264-72.
19. Buee, L., et al., Tau protein isoforms, phosphorylation and role in neurodegenerative disorders. *Brain Res Brain Res Rev*, 2000. **33**(1): 95-130.
20. Olsen, J.V. and M. Mann, Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci USA*, 2004. **101**(37): 13417-22.
21. Domon, B. and R. Aebersold, Mass spectrometry and protein analysis. *Science*, 2006. **312**(5771): 212-7.
22. Witze, E.S., et al., Mapping protein post-translational modifications with mass spectrometry. *Nat Methods*, 2007. **4**(10): 798-806.
23. Perkins, D.N., et al., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 1999. **20**(18): 3551-67.
24. Eng, J.K., A.L. McCormack, and J.R. Yates, 3rd, 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**: 976-989.
25. Ruse, C.I., et al., Motif-specific sampling of phosphoproteomes. *J Proteome Res*, 2008. **7**(5): 2140-50.
26. Beausoleil, S.A., et al., Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci USA*, 2004. **101**(33): 12130-5.
27. Berry, N.B., M. Fan, and K.P. Nephew, Estrogen receptor-alpha hinge-region lysines 302 and 303 regulate receptor

degradation by the proteasome. *Mol Endocrinol*, 2008. **22**(7): 1535-51.
28. Tanner, S., et al., Accurate annotation of peptide modifications through unrestrictive database search. *J Proteome Res*, 2008. **7**(1): 170-181.
29. Martin, D.M., et al., Prophossi: automating expert validation of phosphopeptide-spectrum matches from tandem mass spectrometry. *Bioinformatics*, 2010. **26**(17): 2153-9.
30. Lu, B., et al., Automatic validation of phosphopeptide identifications from tandem mass spectra. *Anal Chem*, 2007. **79**(4): 1301-10.
31. Olsen, J.V., et al., Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 2006. **127**(3): 635-48.
32. Beausoleil, S.A., et al., A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol*, 2006. **24**(10): 1285-92.
33. Ruttenberg, B.E., et al., PhosphoScore: an open-source phosphorylation site assignment tool for MS^n data. *J Proteome Res*, 2008. **7**(7): 3054-9.
34. Villen, J., et al., Large-scale phosphorylation analysis of mouse liver. *Proc Natl Acad Sci USA*, 2007. **104**(5): 1488-93.
35. Manning, G., et al., The protein kinase complement of the human genome. *Science*, 2002. **298**(5600): 1912-34.
36. Obenauer, J.C., L.C. Cantley, and M.B. Yaffe, Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, 2003. **31**(13): 3635-3641.
37. Blom, N., S. Gammeltoft, and S. Brunak, Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol*, 1999. **294**(5): 1351-62.
38. Ahmad, W., et al., Serine 204 phosphorylation and O-beta-GlcNAC interplay of IGFBP-6 as therapeutic indicator to regulate IGF-II functions in viral mediated hepatocellular carcinoma. *Virol J*, 2011. **8**: 208.
39. Xu, H., et al., Toward a complete *in silico*, multi-layered embryonic stem cell regulatory network. *Wiley Interdiscip Rev Syst Biol Med*, 2010. **2**(6): 708-33.
40. Whisenant, T.C., et al., Computational prediction and experimental verification of new MAP kinase docking sites and substrates including Gli transcription factors. *PLoS Comput Biol*, 2010. **6**(8).
41. Iliuk, A.B., et al., In-depth analyses of kinase-dependent tyrosine phosphoproteomes based on metal ion-functionalized soluble nanopolymers. *Mol Cell Proteomics*, 2010. **9**(10): 2162-72.
42. Kim, M.S., et al., Systematic evaluation of alternating CID and ETD fragmentation for phosphorylated peptides. *Proteomics*, 2011. **11**(12): 2568-72.
43. Golub, G.H. and C.F. Van Loan, *Matrix computations*. 3rd ed. 1996, Johns Hopkins University Press.