

## Colander: A Probability-Based Support Vector Machine Algorithm for Automatic Screening for CID Spectra of Phosphopeptides Prior to Database Search

Bingwen Lu,<sup>†</sup> Cristian I. Ruse,<sup>†</sup> and John R. Yates III\*

*Department of Chemical Physiology, SR-11, The Scripps Research Institute, La Jolla, California 92037*

Received February 13, 2008

**Abstract:** We developed a probability-based machine-learning program, Colander, to identify tandem mass spectra that are highly likely to represent phosphopeptides prior to database search. We identified statistically significant diagnostic features of phosphopeptide tandem mass spectra based on ion trap CID MS/MS experiments. Statistics for the features are calculated from 376 validated phosphopeptide spectra and 376 nonphosphopeptide spectra. A probability-based support vector machine (SVM) program, Colander, was then trained on five selected features. Data sets were assembled both from LC/LC-MS/MS analyses of large-scale phosphopeptide enrichments from proteolyzed cells, tissues and synthetic phosphopeptides. These data sets were used to evaluate the capability of Colander to select pS/pT-containing phosphopeptide tandem mass spectra. When applied to unknown tandem mass spectra, Colander can routinely remove 80% of tandem mass spectra while retaining 95% of phosphopeptide tandem mass spectra. The program significantly reduced computational time spent on database search by 60–90%. Furthermore, prefiltering tandem mass spectra representing phosphopeptides can increase the number of phosphopeptide identifications under a predefined false positive rate.

**Keywords:** protein phosphorylation • support vector machine • phosphopeptide • tandem mass spectrometry

### Introduction

Protein phosphorylation is a key post-translational modification (PTMs), regulating many biological processes.<sup>1</sup> To enable broad and large-scale analyses of how changes in phosphorylation states regulate processes in cells, new approaches are under development and refinement. Tandem mass spectrometry is an important analysis tool for post-translational modifications as it is capable of not only identifying the amino acid sequence of a peptide, but also pinpointing the location of a modification within the sequence. Analyses of phosphopeptides using mass spectrometers are frequently performed on peptides obtained by trypsin digestion of proteins. As the pattern of post-translational modifications in a protein can be just as important

as a specific site of modification, methods to analyze larger peptide fragments or even intact proteins are increasingly important.<sup>2</sup> The analysis of phosphorylation has an additional complication of stoichiometry as not all of a specific site may be occupied,<sup>3</sup> thus, favoring the use of enrichment techniques. A variety of enrichment strategies have been developed for improving detection of phosphopeptides. Antibodies are an effective tool for enrichment of phosphotyrosine (pY)-containing peptides<sup>4,5</sup> prior to mass spectrometry. Phosphoserine (pS)/phosphothreonine (pT) containing peptides are usually enriched by stand alone procedures or by a combination of techniques such as immobilized metal ion affinity chromatography (IMAC),<sup>6,7</sup> chemical replacement of the phosphate group with an affinity tag,<sup>8</sup> strong cation exchange (SCX) chromatography,<sup>9</sup> or titanium oxide chromatography.<sup>10</sup>

Collision-induced dissociation (CID) is widely used to fragment both peptides and phosphopeptides to generate the information necessary to identify the amino acid sequence and identify the site of modification. A major challenge to identifying sites of phosphorylation from CID tandem mass spectra is the facile loss of phosphoric acid from phosphoserine and phosphothreonine.<sup>11</sup> This neutral loss can dominate fragmentation of the peptide backbone and yield little or no sequence ions. A neutral-loss triggered MS3 step can be added to enhance fragmentation of the peptide backbone to improve identification<sup>9,12</sup> and quantification<sup>13</sup> of the phosphopeptide. New dissociation methods such as electron capture dissociation (ECD)<sup>14,15</sup> and ETD<sup>16</sup> are capable of generating phosphopeptide spectra without neutral losses and thus generate potentially more informative fragmentation patterns. One study suggests that the higher sensitivity and faster duty cycle of CID-MS/MS could be used for increased phosphopeptide identification by ETD in a nonlinear Paul trap.<sup>17</sup> Alternatively, CID-MS/MS and ECD-MS/MS could be used independently to complement each other and potentially reduce the false positive rates provided postacquisition algorithms (PhosTShunter) are applied.<sup>18</sup> Two post database-search algorithms were developed for phosphorylation site localization,<sup>19,20</sup> both of which utilize the binomial probability models to calculate site localization scores from site-determining ions observed in CID-tandem mass spectra. More recently, Schlosser et al.<sup>21</sup> proposed the use of global neutral loss features for phosphorylation site analysis in individual proteins. Dephosphorylation was also employed for the validation of phosphorylation for QTOF or LTQ-FT data.<sup>22–24</sup> Lu et al.<sup>25</sup> also published two automatic methods for validation of phosphorylation identification from CID-tandem mass spectra.

\* Corresponding author. E-mail: jyates@scripps.edu.

<sup>†</sup> These two authors contributed equally to this work.

**Colander: Probability-Based SVM Algorithm**

In this work, we focus on presearch filtering of pS/pT-containing phosphopeptide spectra from CID-tandem mass spectra.

Rapid scanning mass spectrometers such as a linear ion trap can routinely acquire hundreds of thousands of tandem mass spectra in large-scale proteomic studies. Interrogation of tandem mass spectra to identify modified peptides produces a large computational burden for database search programs.<sup>26</sup> One method to simplify the search process is to establish if a tandem mass spectrum contains the suspected modification before database search. If nonphosphorylated tandem mass spectra can be filtered out before database searching, then this will reduce the number of spectra that have to be searched for modifications.

We describe a probability-based support vector machine (SVM) program to identify tandem mass spectra containing pS/pT. SVM has been used to validate SEQUEST search output,<sup>27</sup> to determine charge states for low-resolution tandem mass spectra,<sup>28</sup> to classify mass spectrometry data of diseased and normal samples,<sup>29</sup> and to validate phosphopeptide/spectrum matches.<sup>25</sup> To develop the algorithm, we identified five features that are related to the labile neutral losses of phosphopeptides during mass spectrometry experiment. Two of the features are related to the neutral losses from the base peak ion. The other three features are related to the neutral losses from any ion. We found that using these features for SVM resulted in good classifying power for phosphopeptide spectra and nonphosphopeptide spectra. The algorithm was evaluated on complex sets of tandem mass spectra of enriched phosphopeptide mixtures, of a nonenriched shotgun peptide mixture, and of synthetic peptides containing phosphoserines.

**Experimental Methods****Preparation of Phosphopeptides from Complex Mixtures.**

Protein concentration was determined using a modified Bradford assay (Bio-Rad). We used proteins from three different sources: nuclear extract from rat brain, whole cell lysates of PC12 and HEK cells. Phosphatase Inhibitor Cocktail (Calbiochem) was used to preserve the phosphorylation of proteins. Nuclear proteins were extracted from clarified rat brain tissue homogenate with a nuclear/cytosolic fractionation kit (BioVision, Inc.). PC12 cells were treated with cAMP for 5 min prior to harvesting and lysis. HEK cells overexpressing beta-1 adrenergic receptor were treated with isoproterenol (Sigma) prior to harvesting and lysis.

One milligram of proteins was further processed by methanol/chloroform extraction. Protein pellet was dissolved by sonication in 50% MeOH in 100 mM Tris-HCl (pH 7.6) and digested with trypsin (1:50) at 37 °C overnight. Phosphopeptides were enriched by a strategy that will be described elsewhere. Enriched phosphopeptides were then subjected to MudPIT analysis.

**Preparation of Casein Proteins.** Casein proteins were obtained from Sigma. Proteins were dissolved in 100 mM Tris (pH 7.6) and digested with trypsin (1:50) at 37 °C overnight. Phosphopeptides were enriched by a strategy that will be described elsewhere. Enriched casein phosphopeptides were then analyzed by LC-MS/MS.

**Yeast Whole Cell Lysate and Digestion.** A protease-deficient *Saccharomyces cerevisiae* strain BJ5460<sup>30</sup> was purchased from American Type Culture Collection (Manassas, VA). The strain was grown to midlog phase (OD 0.6) in YPD, and cells collected by centrifugation were lysed as described previously.<sup>31</sup> The

lysed cells were separated into three fractions (soluble, lightly and heavily washed), and the soluble fraction was used in this study. The soluble fraction of cells was digested by a method slightly modified from the one described previously.<sup>31</sup> Urea was added to the soluble fraction of the cell lysate to denature the proteins. Proteins were then reduced with TCEP (5 mM final concentration) for 20 min, alkylated using iodoacetamide (IAM, 10 mM final concentration) for 15 min in the dark, and subsequently digested with trypsin. The digestion process was stopped by adding formic acid to a final concentration of 1%. The protein digest was aliquoted and stored at -80 °C prior to MudPIT analysis by an LTQ-Orbitrap mass spectrometer.

**MudPIT Analysis.** Samples were pressure-loaded onto a 250- $\mu$ m i.d. fused silica capillary column containing 3 cm of 5- $\mu$ m Aqua C<sub>18</sub> material (Phenomenex, Ventura, CA) followed by 3 cm of 5- $\mu$ m Partisphere strong cation exchanger (Whatman, Clifton, NJ) and capped with a 2  $\mu$ m filtered union. The biphasic column was washed with buffer A. The biphasic column was then connected to an analytical column of a 100- $\mu$ m i.d. capillary with a 5- $\mu$ m pulled tip and packed with 12~13 cm of 3- $\mu$ m Aqua C<sub>18</sub> material (Phenomenex, Ventura, CA).

The column was placed inline with an Agilent 1100 quaternary HPLC and analyzed using a 12-step separation. The buffer solutions were 5% acetonitrile/0.1% formic acid (buffer A), 80% acetonitrile/0.1% formic acid (buffer B), and 500 mM ammonium acetate/5% acetonitrile/0.1% formic acid (buffer C). Step 1 consisted of 15 min of 100% A followed by a gradient of 80 min from 0 to 55% B, reversal to 100% A in 2 and 3 min of re-equilibration with 100% A. Steps 2–11 have the following profile: 3 min of 100% buffer A, 2 min of X% buffer C, a 10 min gradient from 0–15% buffer B, and a 97 min gradient from 15–45% buffer B. The 2 min buffer C percentages (X) are 5, 10, 20, 30, 40, 50, 60, 70, 80, and 90%, respectively, for the 12-step analysis. In the final step, the gradient contained 3 min of 100% buffer A, 20 min of 100% buffer C, a 10 min gradient from 0–15% buffer B, and a 107 min gradient from 15–70% buffer B.

Peptides were electrosprayed directly into an LTQ mass spectrometer (nuclear extract from rat brain, PC12 cells) or an LTQ-Orbitrap mass spectrometer (HEK cells and yeast whole cell lysates) (ThermoFinnigan, Palo Alto, CA). A cycle consisted of full scan MS (400–1800 *m/z*) followed by MS/MS on the seven most abundant ions. Normalized collision energy of 35% and an isolation width of 3 *m/z* units were used for acquisition of tandem mass spectra.

**LC-MS/MS Analysis.** Digested casein proteins were pressure-loaded onto 100- $\mu$ m i.d. fused silica capillary with a 5- $\mu$ m pulled tip packed with 10 cm of 3- $\mu$ m Aqua C18. After loading, the column was placed inline with an Agilent 1100 quaternary HPLC and analyzed using the following gradient: 2 min of 100% buffer A, a 13 min gradient from 0 to 10% buffer B, a 40 min gradient from 10 to 55% buffer B, a 10 min gradient from 55% to 100% buffer B, 5 min of 100% B, reversal to 100% A in 2 and 8 min of re-equilibration.

**Analysis of Tandem Mass Spectrometry Data.** Raw files were extracted with an in-house software<sup>32</sup> to produce distinct sets of MS2 files containing tandem mass spectra. When tandem mass spectra (rat brain, PC12 cells, and casein proteins) were collected from an LTQ mass spectrometer, the data were searched using SEQUEST<sup>33</sup> with differential modification of +80 Da on STY (phosphorylation),  $\pm 3$  Da for precursor mass tolerance, with or without enzyme restriction, as indicated in Results and Discussion. When tandem mass spectra (HEK cells

**Table 1.** Some Statistics of Extracted Features for the Positive and Negative Training Set<sup>a</sup>

		peak-NL pairs (98)	NL/BP ratio	peak-NL pairs/2+	peak-NL pairs (80)	H <sub>2</sub> O loss/BP ratio	double NL/BP ratio <sup>a</sup>
Positive	Mean	16.58511	0.671794	12.5984	13.94947	0.020082	0.051537
	STDEV	3.368697	0.378049	2.99991	2.937876	0.096731	0.108425
Negative	Mean	14.13032	0.028291	11.26596	12.85638	0.311363	0.010688
	STDEV	2.66544	0.106396	2.820316	2.722487	0.339801	0.046665
	FCS	0.998648	0.854783	0.305042	0.211088	0.19436	0.010759

<sup>a</sup> The feature "Double NL/BP Ratio" was not used in the final classifier.

and yeast whole cell lysates) were collected from an LTQ-Orbitrap mass spectrometer, the data were searched using SEQUEST with differential modification of +79.966332 Da on STY, ±50 ppm for precursor mass tolerance, with or without enzyme restriction. The rat brain and PC12 cells data were searched against the EBI-IPI rat protein database (version 3.05, April 2005) with its reverse decoy. The HEK cells data were searched against the EBI-IPI human database (version 3.04, March 2005) with its reverse decoy. The yeast whole cell lysates data were searched against the Saccharomyces Genome Database (SGD) protein database (version December 2005) with its reverse decoy. The casein proteins data were searched against a database containing all caseins ( $\alpha$ -s1,  $\alpha$ -s2,  $\beta$  and  $\kappa$ ), the whole SGD proteins, and standard contaminant proteins with its reverse decoy.

SEQUEST results were analyzed and filtered with the following options using DTASelect2.0:<sup>34</sup> --fp 0.05, selects only peptide/spectrum matches with a false positive rate of 5% at spectrum level; -p 1, considers proteins identified by a single peptides; -m 0, display (for visualization) only the modified peptides.

**Obtaining Probability Estimates from Support Vector Machines (SVMs).** Support vector machine (SVM) is one of the supervised statistical learning methods for classification problems, which can also be applied to regression and ranking problems.<sup>35–37</sup> For a two-class classification problem, the SVM classifier will find a hyperplane to separate the data into two classes by implementing the following ideas.

First, the SVM classifier searches for a hyperplane that separates the two classes with a maximum margin. Second, for data sets that could not be separated by a simple hyperplane, SVM maps the input vectors into a high dimensional feature space and constructs the Optimal Separating Hyperplane (OSH) in the feature space, a process known as kernel mapping. Finally, for data containing noisy or mislabeled data points, SVM introduces a slack parameter that controls the tradeoff between margin and misclassification error. The optimized hyperplane found by SVM may nearly, but not perfectly, separate the two classes, allowing a few data points to be misclassified.

For a given test example  $x$ , an SVM classifier outputs a predictive score that provides the distance of  $x$  from the optimal separating hyperplane in the feature space. This sign of this predictive score indicates to which class  $j$  example  $x$  belongs, where  $j \in \{+1, -1\}$ . However, knowing the class label (+, or -) or the predictive value many times is not good enough to evaluate a classification. A better choice is to convert the predictive values to posterior probability estimates. We used a technique known as binning to convert predictive values to posterior probabilities,<sup>38</sup> which was implemented internally in the LIB-SVM software package.<sup>39</sup>

## Results and Discussion

**Phosphopeptide Spectra and Nonphosphopeptide Spectra Data Sets.** Known sets of 376 phosphopeptide and 376 non-phosphopeptide tandem mass spectra from a previous study<sup>25</sup> were used as training data sets for the probability-based SVM program Colander. Known sets of 308 phosphopeptide tandem mass spectra and 1011 nonphosphopeptide tandem mass spectra<sup>25</sup> were used as a testing set. Various tandem mass spectra were used to evaluate the performance of the program, including (1) 124 545 tandem mass spectra from a MudPIT experiment of phosphopeptides enriched from a nuclear extract of rat brain on a linear ion trap mass spectrometer; (2) 319 728 tandem mass spectra from a MudPIT experiment of phosphopeptides enriched from a nuclear extract of PC12 cells on a linear ion trap mass spectrometer; (3) 132 011 tandem mass spectra from a MudPIT experiment of phosphopeptides enriched from nuclear extract of HEK cells on an LTQ-Orbitrap mass spectrometer; (4) 81 112 tandem mass spectra from a MudPIT experiment of soluble fraction of whole cell lysate of *S. cerevisiae* on an LTQ-Orbitrap mass spectrometer; (5) 248 714 tandem mass spectra from a MudPIT experiment of phosphopeptides enriched from casein proteins on a linear ion trap mass spectrometer. A total of 281 tandem mass spectra collected from the direct injection of synthetic phosphopeptides into the LTQ mass spectrometer as described previously<sup>25</sup> were also used to validate the program.

**Characterizing Training Data Sets.** CID MS/MS spectra are known to show neutral losses from both the precursor ions and fragment b-ions and y-ions.<sup>18,25</sup> For each spectrum/peptide match, we extracted a 6-dimensional neutral loss related feature vector  $\bar{x} = (f_1, f_2, \dots, f_6)$  in  $R^6$  where  $f_i$  is the  $i$ -th feature below:

- (1) Number of peaks-“neutral loss (NL) of 98” pairs;
- (2) Precursor “neutral loss (NL) of 98”/“base peak (BP)” intensity ratio;
- (3) Number of peaks-“NL of 49” pairs (the fragment ion is of charge +2);
- (4) Number of peaks-“NL of 80” pairs;
- (5) Precursor “loss of water (18 Da)”/“base peak (BP)” intensity ratio;
- (6) Double precursor “neutral loss (NL) of 98”/“base peak (BP)” intensity ratio.

The mean and standard deviation of each feature were calculated for the positive training set and negative training set (see Table 1 for more details). The distributions for the positive set and negative set were analyzed and the features that differ significantly between the positive set and negative set are selected for discussion below. All these features are associated with the neutral losses of phosphate groups or water.

The first three features are well-known features for tandem mass spectra of phosphopeptides. The first feature is the presence of large number of ions associated with a neutral loss of 98. This feature is associated with the loss of H<sub>3</sub>PO<sub>4</sub> from

**Colander: Probability-Based SVM Algorithm**

both the precursor ion and fragment ions (b-ions and y-ions). The second feature is the intensity of the precursor ion neutral loss. Frequently, the precursor ion neutral loss is also the base peak. The third feature is essentially the same as the first feature, only that the peak and the corresponding neutral loss ion are both doubly charge.

The fourth feature is associated with the loss of  $\text{HPO}_3$  from both precursor and fragment ions. Phosphopeptides sometimes lose the  $\text{HPO}_3$  group during mass spectrometry experiments.<sup>11</sup> The fifth feature is related to the loss of water ( $\text{H}_2\text{O}$ ) from the precursor ion. Our study showed that it is less likely for phosphopeptides to lose water only, without losing the phosphate group. On average, the intensity of water loss peaks is 2% of the base peak in phosphopeptide tandem mass spectra. However, the average intensity of water loss peaks is 31% of the base peak in nonphosphopeptide tandem mass spectra. Finally, the last feature is associated with double neutral losses from precursor ions. Multiply phosphorylated peptides frequently lose more than one phosphate group from a precursor ion.

To see which features might provide the most discriminatory power for our following analysis using SVM, the Fisher criterion score (FCS) for each feature was computed. For a pair of distributions A and B for a specific feature, with means  $\mu_A$  and  $\mu_B$  and standard deviations  $\sigma_A$  and  $\sigma_B$ , the FCS is defined as

$$\text{FCS} = \frac{(\mu_A - \mu_B)^2}{\sigma_A + \sigma_B} \quad (1)$$

The higher the FCS score for a feature, the greater the difference between the positive set and the negative set will be for that feature. The FCS scores for the selected six features are given in Table 1. The sixth feature, double precursor “neutral loss (NL) of 98”/“base peak (BP)” ratio, has the smallest FCS score. SVM classification without this feature showed that there was no loss of classification accuracy. However, omission of any of the first five features led to decreased classification accuracy by SVM. As a consequence, only the first five features were used for the following SVM analysis.

**Classifications by SVM.** We used a binary classifier support vector machine (SVM) to classify CID tandem mass spectra. Each tandem mass spectrum was classified as one of the two classes: phosphopeptide spectrum (+), or nonphosphopeptide spectrum (-). The SVM classifier also assessed the probability of being a phosphopeptide spectrum for each tandem mass spectrum. A tandem mass spectrum will be classified as phosphopeptide spectrum (+) if its probability score is greater than 0.5. A tandem mass spectrum will be classified as nonphosphopeptide spectrum (-) if its probability score is less than 0.5. A tandem mass spectrum will remain unclassified if the probability score equals 0.5, which is a relatively rare situation from our observation.

We evaluated the classification power of SVM using the five selected features described above. We used the LIB-SVM package and trained the program using 376 positive (phosphorylated) spectra and 376 negative (nonphosphorylated) spectra. The following kernel functions were tried: linear kernel function, polynomial kernel function with  $d = 2$ , and radial basis function (RBF). The prediction accuracies on the testing sets are shown in Table 2 and the detailed performance of the three kernel functions are shown in Supporting Table 1. The data showed that the SVM classifier with the RBF kernel has the best power to retain phosphopeptide spectra (94.16%) while removing the most nonphosphopeptide spectra (94.56% = 1 - 5.44%). To evaluate if the differences for the different kernel

**Table 2.** Classification Results on the Testing Data

kernel	positive testing	negative testing
Linear	73.38%	1.38%
Polynomial ( $d = 2$ )	83.77%	5.24%
RBF	94.16%	5.44%

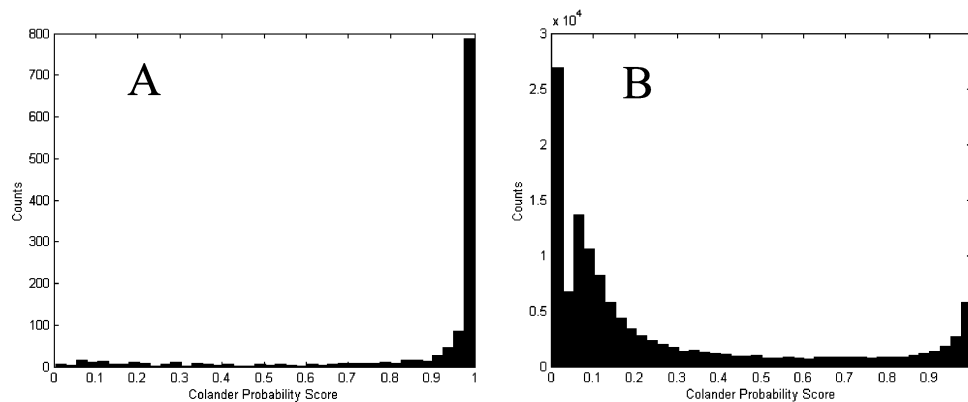
methods are significant, we performed the Fisher's exact test. The  $p$ -value for the classification difference on the test data between the linear kernel and RBF kernel is  $5.4 \times 10^{-5}$ , and that for the classification difference between the polynomial kernel (with  $d = 2$ ) and RBF kernel is 0.019. Thus, the RBF kernel showed statistically better classification power than the linear kernel and the polynomial kernel ( $d = 2$ ). The RBF kernel was employed in the final Colander program that was used for the following analysis.

**Distribution of Colander Probability Scores.** We show in Figure 1 the distribution of Colander probability scores for tandem mass spectra matched to phosphopeptides by SEQUEST (Figure 1A) and the distribution of Colander probability scores not matched to phosphopeptides (Figure 1B), using our tandem mass spectrometry data collected from nuclear extracts of rat brain. From this figure, we see that a majority of spectra that are matched to phosphopeptides show high Colander probability scores (close to 1). However, for the spectra that do not match to phosphopeptides, a large portion of spectra show Colander probability scores close to 0. The distributions of other data sets from phosphopeptide enrichment experiments showed a similar pattern. This confirms that we can use Colander probability scores to filter tandem mass spectra for the presence of phosphorylation.

**Filtering Performance of Colander on Tandem Mass Spectra from Phosphopeptide Enrichment Experiments.** To evaluate the efficacy of using Colander to identify phosphopeptide containing tandem mass spectra, we applied the program to three different MS/MS data sets generated from phosphopeptide enrichment experiments.

The first MS/MS data set was collected in a MudPIT experiment of phosphopeptides enriched from a nuclear extract of rat brain on a linear ion trap mass spectrometer. There are 124 545 tandem mass spectra for this data set. A SEQUEST database search with no enzyme restriction followed by DTA-Select filtering identified 484 phosphopeptides corresponding to 1204 phosphopeptide tandem mass spectra with a 5% false positive rate at the spectrum level. Applying Colander to this data set at a 0.5 probability cutoff, we obtained 25 876 tandem mass spectra (20.8% of the original spectra). SEQUEST search and DTASelect filtering with the same parameters identified 537 phosphopeptides corresponding to 1269 phosphopeptide spectra, again with a 5% false positive rate at the spectrum level (Table 3). We identified more peptides and spectra at a predefined false positive rate of 5%. The same observation was also made for the other data sets as shown below. A discussion will then be given on why we can identify more peptides and spectra at a predefined false positive rate afterward.

The second data set is collected in a MudPIT experiment of phosphopeptides enriched from nuclear extract of PC12 cells, again on a linear ion trap mass spectrometer. This data set has 319 728 tandem mass spectra. SEQUEST database search with trypsin digestion restriction followed by DTASelect filtering identified 2073 phosphopeptides corresponding to 11 506 phosphopeptide spectra. If we apply Colander filtering to the 319 728 tandem mass spectra and use a probability cutoff of



**Figure 1.** Distribution of Colander probability scores. (A) Distribution of Colander probability scores for the spectra matched to phosphopeptides by SEQUEST (after DTASelect filtering). (B) Distribution of probability scores for tandem mass spectra with phosphopeptide spectra subtracted.

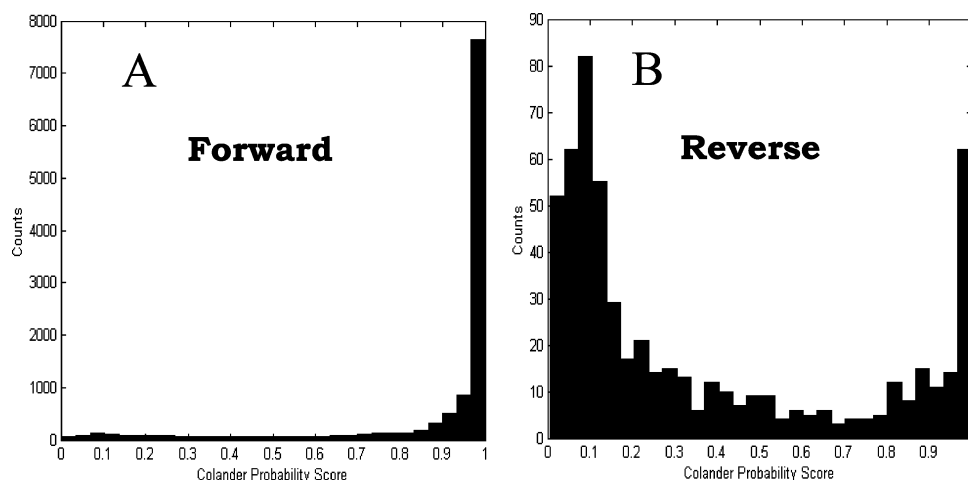
**Table 3.** Filtering Performance of Colander on MS/MS Data from Phosphopeptide Enrichment Experiments

		no.spectra	% spectra	no. phosphospectra	% phosphospectra
Rat Brain	No-Filtering	124545	100.00%	1204	100.00%
	Colander	25876	20.78%	1269	105.40%
PC12 Cells	No-Filtering	319728	100.00%	11496	100.00%
	Colander	68158	21.32%	11645	101.30%
HEK Cells	No-Filtering	132011	100.00%	414	100.00%
	Colander	34546	26.17%	480	115.94%

0.5, we obtain 68 158 tandem mass spectra expected to contain phosphopeptides (21.3% of original spectra). SEQUEST search and DTASelect filtering with the same parameters identified 2141 phosphopeptides corresponding to 11 656 phosphopeptide tandem mass spectra (Table 3).

The third data set is collected in a MudPIT experiment of phosphopeptides enriched from nuclear extract of HEK cells on an LTQ-Orbitrap mass spectrometer. This data set has 132 011 tandem mass spectra. SEQUEST database search with trypsin enzyme restriction followed by DTASelect filtering identified 185 phosphopeptides corresponding to 414 phosphopeptide tandem mass spectra. With Colander filtering at a probability cutoff of 0.5, we obtained 34 546 tandem mass spectra (26.2% of original spectra). SEQUEST search and DTASelect filtering with the same parameters identified 225 phosphopeptides corresponding to 480 phosphopeptide spectra (Table 3).

We observed that frequently, under a predefined false positive rate, the number of phosphopeptides identified from the smaller MS/MS data sets with Colander filtering is greater than the number of phosphopeptides identified from the bigger MS/MS data sets without Colander filtering. We suspected that Colander might remove tandem mass spectra that can be matched to reversed phosphopeptide sequences. To further examine this issue, we plotted the distribution of Colander scores for the PC12 spectra matched to forward phosphopeptides and reverse phosphopeptides in Figure 2, panels A and B, respectively. The plots show that, although the majority of tandem mass spectra matched to forward phosphopeptides are of high Colander probability scores, a large portion of tandem mass spectra matched to reverse phosphopeptides are of Colander probability scores that are lower than the normal filtering probability cutoff of 0.5. Actually, the probability



**Figure 2.** Distributions of Colander probability scores for spectra matched to forward phosphopeptides and reverse phosphopeptides. (A) Distribution of scores for spectra matched to forward phosphopeptides; (B) distribution of scores for spectra matched to reverse phosphopeptides.

**Colander: Probability-Based SVM Algorithm**

distribution of the tandem mass spectra matched to the reverse phosphopeptides (Figure 2B) is very similar to the probability distribution of the tandem mass spectra not matched to phosphopeptides (Figure 1B). The plots confirmed our suspicion and showed that Colander can effectively remove a large portion of tandem mass spectra that will match to reverse phosphopeptides. The removal of these tandem mass spectra will affect the postsearch filtering of DTASelect. DTASelect will lower the filtering threshold to achieved the same false positive rate (5% in this case), finally leading to more spectrum and peptide identifications, as shown by the data in Table 3. Filtering details of tandem mass spectra by Colander on these three data sets using different probability thresholds are given in Supporting Figure 1.

**Filtering Performance of Colander on Tandem Mass Spectra from Nonenrichment Experiment.** Colander shows its power in filtering out nonphosphopeptide spectra when applied to tandem mass spectra collected from peptide mixtures not enriched for phosphopeptides. When applied to 81 112 tandem mass spectra from a MudPIT experiment of soluble fraction of a whole cell lysate of *S. cerevisiae* on an LTQ-Orbitrap mass spectrometer at a threshold value of 0.5, Colander retained 94.74% (18 out of 19) of phosphopeptide tandem mass spectra while only retaining 7.32% of nonphosphopeptide spectra. When the filtering threshold was increased to 0.95, Colander kept 89.47% (17 out of 19) of phosphopeptide spectra while removing 99.01% of total spectra. A closer examination of the phosphopeptide tandem mass spectra that did not pass the Colander threshold value of 0.5 showed that this spectrum of charge +2 was matched to a phosphopeptide with a borderline SEQUEST XCorr score of 2.0872. Filtering details of tandem mass spectra by Colander using different probability thresholds are given in Supplementary Table 1. Similar results were observed from MS/MS data from other nonenrichment experiments.

We suggest that Colander might function as an *in silico* enrichment approach for the identification of phosphopeptides from tandem mass spectra generated from nonenriched samples. Specifically, parallel identification can be carried out for phosphopeptides and unmodified peptides. Database searching for nonphosphorylated peptides can be carried out on all acquired MS/MS data set. At the same time, tandem mass spectra can be subjected to Colander filtering to generate a smaller set of spectra for phosphorylation interrogation. While retaining the majority of phosphorylation matches, one can save more than 90% of computational time.

**Capability of Colander To Retain Spectra of Peptides with Multiple Phosphorylation Sites.** We also tested the capability of Colander to retain spectra that are generated from peptides containing multiple phosphorylation sites. Tandem mass spectra generated from the casein experiment were used for this purpose. For this data set, there are 682, 1794, 324, and 411 redundant phosphopeptide spectra that are matched to casein phosphopeptides, corresponding to peptides with 1, 2, 3, and 4 phosphorylation sites, respectively (Table 4). At the filtering threshold of probability 0.5, Colander is able to retain 94.13%, 98.10%, 98.46%, and 94.65% for spectra identified to peptides with 1, 2, 3, and 4 phosphorylation sites, respectively. The results show that Colander retains spectra of peptides with multiple phosphorylation sites. Distributions of the Colander probability scores are shown in Supporting Figure 2.

**Validation of the Method with Synthetic Peptides.** We further validated the program using spectra generated from

**Table 4.** Capability of Colander To Retain Spectra of Multiply Phosphorylated Peptides

no. phosphosites	no. spectra	no. spectra passed filter	% retained
1	682	642	94.13%
2	1794	1760	98.10%
3	324	319	98.46%
4	411	389	94.65%

the synthetic phosphopeptides pSFVLNPTNIGMSKSSQGH-VTK and SFVLNPTNIGMpSKSSQGHVTK.<sup>25</sup> Sites of phosphorylation for these two synthetic peptides are indicated by the lowercase "p" as in "pS". A total of 281 redundant tandem mass spectra were acquired on a linear ion trap mass spectrometer. These 281 spectra were subjected to the filtering by Colander and all of them passed the filtering threshold of probability 0.5. Actually, all of them will pass filtering even if we increase the threshold to 0.95. The distribution of the Colander probability scores for these spectra is shown in Supporting Figure 3.

**Computational Time and Availability.** Another consideration of the performance of Colander is computation time. For the 124 545 tandem mass spectra from nuclear extract of rat brain, it took Colander 1963 s (or 0.5 h) to carry out the filtering step. The computational time spent on the phosphorylation search with no enzyme restriction on the same data set is 18 677 597 s (or 5188.2 h, roughly 7 months by a single-CPU computer). The time spent on filtering step is only 0.01% of the time spent on database search. For comparison purposes, the computational time spent on the phosphorylation search with no enzyme restriction on the data after Colander filtering is 6 087 498 s (or 1691.0 h, roughly 70 days), 32.6% of the time spent on the whole data set (Supporting Table 2).

For another data set, the 319 728 tandem mass spectra from nuclear extract of PC12 cells, it took Colander 1983 s to filter out nonphosphopeptide spectra. The computational time spent on the phosphorylation search with trypsin restriction on the same data set is 1 331 596 (or 369.9 h). The time spent on the filtering step is only 0.15% of the time on the database search. The time spent on the phosphorylation search with trypsin restriction on the data after Colander filtering is 335 834 s (or 93.3 h), 25.2% of the time spent on the whole data set (Supporting Table 2). The computational time for HEK cells data, yeast lysates data, and casein proteins data can be found in Supporting Table 2 as well.

In summary, the computational time spent on the presearch filtering of Colander is negligible compared to the time spent on database search. Meanwhile, filtering by Colander can significantly reduce computational time required for database search.

The Colander program and the testing data set of 308 phosphopeptide MS/MS spectra and 1011 nonphosphopeptide MS/MS spectra in MS2 format are available for academic use without charge at <http://fields.scripps.edu/download.php>.

## Conclusions

For mass spectrometry analysis of phosphopeptides, a majority of tandem mass spectra are not generated from phosphopeptides. As a consequence, the majority of computational time is wasted on the interrogation of nonphosphopeptide spectra for the identification of phosphopeptides. We developed a support machine (SVM) algorithm to effectively select phosphopeptide spectra-tandem mass spectra that are

generated from phosphopeptides. The algorithm was trained by using known phosphopeptide spectra and nonphosphopeptide spectra. The algorithm has been implemented in the software package “Colander”. The Colander program can report a phosphopeptide spectrum probability score for each tandem mass spectrum. The probability scores are then used to filter out nonphosphopeptide spectra.

The Colander program can remove nonphosphopeptide spectra, thus, maximizing computational resources. The output of probability scores also gives the users the flexibility to control the filtering threshold. When applied to unknown tandem mass spectra, the algorithm can routinely remove 80% of total tandem mass spectra while keeping more than 95% of phosphopeptide tandem mass spectra. The program is especially useful when applied to tandem mass spectra generated from complex protein mixtures that do not have enrichment for phosphopeptides, where even a larger portion (larger than 99%) of tandem mass spectra are nonphosphopeptide spectra. The Colander program can effectively remove tandem mass spectra that could be matched to reverse phosphopeptides (false positives) and frequently leads to more identifications under a fixed false positive rate. The computational cost by Colander is relatively small (usually less than 0.2%) comparing to the computational time spent on database search. The Colander program is also effective at retaining phosphopeptide spectra generated from phosphopeptides containing multiple phosphorylation sites.

**Abbreviations:** NL, neutral loss; BP, base peak; SVM, support vector machine; MS, mass spectrometry; MS/MS, tandem MS; MudPIT, multiple dimensional protein identification technology; CID, collision induced dissociation; PTM, post-translational modification; pS, phosphoserine; pT, phosphothreonine; pY, phosphotyrosine.

**Acknowledgment.** B.L. is supported by a CFFT computational fellowship BALCH05X5. C.I.R. and J.R.Y. acknowledge support from NIH 5R01MH067880-02 and NIH P41 RR011823-10. The authors thank Dr. Akira Motoyama for providing the yeast whole cell lysate data. The authors thank Drs. Lujian Liao, Aleksey Nakorchevskiy, and Tao Xu for helpful review of the manuscript.

**Supporting Information Available:** Figures of filtering performance of Colander using different filtering threshold on rat brain data (S-Figure 1A), filtering performance of Colander using different filtering threshold on PC12 cells data (S-Figure 1B), filtering performance of Colander using different filtering threshold on HEK cells data (S-Figure 1C), capability of Colander to retain spectra of multiply phosphorylated peptides (S-Figure 2), distribution of probability scores for spectra generated from synthetic phosphoserine containing peptides (S-Figure 3); Tables of performance of the three kernel function on the testing data (S-Table 1), filtering performance of Colander on nonenrichment data (S-Table 2), computational time (S-Table 3). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

(1) Johnson, S. A.; Hunter, T. *Nat. Methods* **2005**, *2*, 17–25.  
 (2) Siuti, N.; Kelleher, N. L. *Nat. Methods* **2007**, *4*, 817–821.  
 (3) Kjeldsen, F.; Savitski, M. M.; Nielsen, M. L.; Shi, L.; Zubarev, R. A. *Analyst* **2007**, *132*, 768–776.

(4) Steen, H.; Kuster, B.; Fernandez, M.; Pandey, A.; Mann, M. *J. Biol. Chem.* **2002**, *277*, 1031–1039.  
 (5) Rush, J.; Moritz, A.; Lee, K. A.; Guo, A.; Goss, V. L.; Spek, E. J.; Zhang, H.; Zha, X. M.; Polakiewicz, R. D.; Comb, M. *J. Nat. Biotechnol.* **2005**, *23*, 94–101.  
 (6) Ficarro, S. B.; McClelland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. *Nat. Biotechnol.* **2002**, *20*, 301–305.  
 (7) Brill, L. M.; Salomon, A. R.; Ficarro, S. B.; Mukherji, M.; Stettler-Gill, M.; Peters, E. C. *Anal. Chem.* **2004**, *76*, 2763–2772.  
 (8) Oda, Y.; Nagasu, T.; Chait, B. T. *Nat. Biotechnol.* **2001**, *19*, 379–382.  
 (9) Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villen, J.; Li, J.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12130–12135.  
 (10) Pinkse, M. W.; Uitto, P. M.; Hilhorst, M. J.; Ooms, B.; Heck, A. J. *Anal. Chem.* **2004**, *76*, 3935–3943.  
 (11) DeGnore, J. P.; Qin, J. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 1175–1188.  
 (12) Ulintz, P. J.; Bodenmiller, B.; Andrews, P. C.; Aebersold, R.; Nesvizhskii, A. I. *Mol. Cell. Proteomics* **2008**, *7*, 71–87.  
 (13) Wolschin, F.; Lehmann, U.; Glinzki, M.; Weckwerth, W. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3626–3632.  
 (14) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. *J. Am. Chem. Soc.* **1998**, *120*, 3265–3266.  
 (15) Cooper, H. J.; Hakansson, K.; Marshall, A. G. *Mass Spectrom. Rev.* **2005**, *24*, 201–222.  
 (16) Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528–9533.  
 (17) Hartmer, R.; Lubeck, M.; Baessmann, C.; Brekenfeld, A. *54th ASMS Conference on Mass Spectrometry*, 2006.  
 (18) Kocher, T.; Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. *J. Proteome Res.* **2006**, *5*, 659–668.  
 (19) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. *Nat. Biotechnol.* **2006**, *24*, 1285–1292.  
 (20) Olsen, J. V.; Blagoev, B.; Gnäd, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. *Cell* **2006**, *127*, 635–648.  
 (21) Schlosser, A.; Vanselow, J. T.; Kramer, A. *Anal. Chem.* **2007**, *79*, 7439–7449.  
 (22) Wu, H.Y.; Tseng, V. S.; Liao, P. C. *J. Proteome Res* **2007**, *6*, 1812–1821.  
 (23) Imanishi, S. Y.; Kochin, V.; Ferraris, S. E.; de Thonel, A.; Pallari, H. M.; Corthals, G. L.; Eriksson, J. E. *Mol. Cell. Proteomics* **2007**, *6*, 1380–1391.  
 (24) Ishihama, Y.; Wei, F. Y.; Aoshima, K.; Sato, T.; Kuromitsu, J.; Oda, Y. *J. Proteome Res.* **2007**, *6*, 1139–1144.  
 (25) Lu, B.; Ruse, C.; Xu, T.; Park, S. K.; Yates, J. R. *Anal. Chem.* **2004**, *79*, 1301–1310.  
 (26) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. *Anal. Chem.* **2005**, *77*, 4626–4639.  
 (27) Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. *J. Proteome Res.* **2003**, *2*, 137–146.  
 (28) Klammer, A. A.; Wu, C. C.; Maccoss, M. J.; Noble, W. S. *Proc. IEEE Comput. Syst. Bioinform. Conf.* **2005**, 175–185.  
 (29) Zhang, X.; Lu, X.; Shi, Q.; Xu, X. Q.; Leung, H. C.; Harris, L. N.; Iglehart, J. D.; Miron, A.; Liu, J. S.; Wong, W. H. *BMC Bioinf.* **2006**, *7*, 197.  
 (30) Jones, E. W. *Methods Enzymol.* **1991**, *194*, 428–453.  
 (31) Washburn, M. P.; Wolters, D.; and Yates, J. R. III. *Nat. Biotechnol.* **2001**, *19*, 242–247.  
 (32) McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J. R.; Cociorva, D.; Yates, J. R., III *Rapid Commun. Mass Spectrom.* **2004**, *18*, 2162–2168.  
 (33) Eng, J.; McCormack, A.; Yates, J., III *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.  
 (34) Cociorva, D.; Yates, J. R., III. DTASelect 2.0: Improving the Confidence of Peptide and Protein Identifications. *Proceedings of the 54<sup>th</sup> ASMS Annual Meeting*, 2006.  
 (35) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Press: New York, 1995.  
 (36) Vapnik, V. *Statistical Learning Theory*; Wiley Press: New York, 1998.  
 (37) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: New York, 2000.  
 (38) Kwok, J. T. Y. *IEEE Trans. Neural Networks* **1999**, *5*, 1018–1031.  
 (39) Chang, C. C.; Lin, C. J. LIBSVM: a library for support vector machines Web page. National Taiwan University, 2001; software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

PR8001194