

# Automatic Validation of Phosphopeptide Identifications from Tandem Mass Spectra

Bingwen Lu, Cristian Ruse, Tao Xu, Sung Kyu Park, and John Yates III\*

Department of Cell Biology, The Scripps Research Institute, La Jolla, California 92037

We developed and compared two approaches for automated validation of phosphopeptide tandem mass spectra identified using database searching algorithms. Phosphopeptide identifications were obtained through SEQUEST searches of a protein database appended with its decoy (reversed sequences). Statistical evaluation and iterative searches were employed to create a high-quality data set of phosphopeptides. Automation of postsearch validation was approached by two different strategies. By using statistical multiple testing, we calculate a *p* value for each tentative peptide phosphorylation. In a second method, we use a support vector machine (SVM; a machine learning algorithm) binary classifier to predict whether a tentative peptide phosphorylation is true. We show good agreement (85%) between postsearch validation of phosphopeptide/spectrum matches by multiple testing and that from support vector machines. Automatic methods conform very well with manual expert validation in a blinded test. Additionally, the algorithms were tested on the identification of synthetic phosphopeptides. We show that phosphate neutral losses in tandem mass spectra can be used to assess the correctness of phosphopeptide/spectrum matches. An SVM classifier with a radial basis function provided classification accuracy from 95.7% to 96.8% of the positive data set, depending on search algorithm used. Establishing the efficacy of an identification is a necessary step for further postsearch interrogation of the spectra for complete localization of phosphorylation sites. Our current implementation performs validation of phosphoserine/phosphothreonine-containing peptides having one or two phosphorylation sites from data gathered on an ion trap mass spectrometer. The SVM-based algorithm has been implemented in the software package DeBunker. We illustrate the application of the SVM-based software DeBunker on a large phosphorylation data set.

Reversible phosphorylation on proteins is one of the most important regulatory post-translational modifications (PTMs), with close to a third of proteins in higher organisms being phosphorylated.<sup>1</sup> Regulation through protein phosphorylation modulates processes such as the cell cycle, cell differentiation, metabolism, protein localization, protein–protein interaction, etc. Abnormal

levels of protein phosphorylation have been detected in both biological models of diseases and tissues from patients with cancer, diabetes, Alzheimer's disease, and heart failure, as well as other major human diseases. Therefore, large-scale identification of protein phosphorylation may pinpoint pathways or processes activated or perturbed as a function of disease.

Most approaches for the identification of phosphorylation sites involve digesting a protein or protein mixture. As phosphorylation is usually present at low stoichiometry, a variety of enrichment strategies have been developed to improve the detection of phosphopeptides. Antibodies are an effective tool for enrichment of phosphotyrosine (pY)-containing peptides<sup>2,3</sup> prior to mass spectrometry. Phosphoserine (pS)/phosphothreonine (pT)-containing peptides are usually enriched by stand-alone procedures or a combination of the following techniques: immobilized metal ion affinity chromatography (IMAC),<sup>4,5</sup> chemical replacement of the phosphate group with an affinity tag such as biotin,<sup>6</sup> strong cation exchange chromatography,<sup>7</sup> and titanium oxide chromatography.<sup>8</sup>

Affinity-based methods are effective for enrichment of phosphopeptides, and tandem mass spectrometers are well suited to identify both the amino acid sequence of the peptide and the site of phosphorylation.<sup>1,9–12</sup> Collision-activated dissociation is widely used to fragment both peptides and phosphopeptides to generate the information necessary to identify the amino acid sequence. A major obstacle in identifying phosphopeptides from CAD-MS2 spectra is the facile loss of phosphate groups from phosphoserine and phosphothreonine.<sup>13</sup> A large percentage of phosphopeptides undergo a significant neutral loss of phosphoric acid, a process

- (2) Steen, H.; Kuster, B.; Fernandez, M.; Pandey, A.; Mann, M. *J. Biol. Chem.* **2002**, *277*, 1031–1039.
- (3) 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.
- (4) 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.
- (5) Brill, L. M.; Salomon, A. R.; Ficarro, S. B.; Mukherji, M.; Stettler-Gill, M.; Peters, E. C. *Anal. Chem.* **2004**, *76*, 2763–2772.
- (6) Oda, Y.; Nagasu, T.; Chait, B. T. *Nat. Biotechnol.* **2001**, *19*, 379–382.
- (7) 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.
- (8) Pinkse, M. W.; Uitto, P. M.; Hilhorst, M. J.; Ooms, B.; Heck, A. J. *Anal. Chem.* **2004**, *76*, 3935–3943.
- (9) Mann, M.; Jensen, O. N. *Nat. Biotechnol.* **2003**, *21*, 255–261.
- (10) Cantin, G. T.; Yates, J. R., III. *J. Chromatogr., A* **2004**, *1053*, 7–14.
- (11) Garcia, B. A.; Shabanowitz, J.; Hunt, D. F. *Methods* **2005**, *35*, 256–264.
- (12) Jensen, O. N. *Curr. Opin. Chem. Biol.* **2004**, *8*, 33–41.
- (13) DeGnove, J. P.; Qin, J. J. *Am. Soc. Mass Spectrom.* **1998**, *9*, 1175–1188.

\* To whom correspondence should be addressed. E-mail: jyates@scripps.edu.

(1) Johnson, S. A.; Hunter, T. *Nat. Methods* **2005**, *2*, 17–25.

known as  $\beta$ -elimination, where pS and pT convert to  $\beta$ -dehydroalanyl and dehydroamino-2-butyrate, respectively. Consequently, fragmentation of the peptide backbone yields few or no sequence ions. To improve the fragmentation information of peptides undergoing neutral loss, this feature in a spectrum can be used to trigger an MS3 event to fragment the neutral loss ion. More efficient fragmentation of the peptide backbone is observed, allowing higher quality identification of the phosphopeptide.<sup>7</sup> 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 potentially more informative fragmentation patterns. A recent study suggests that the higher sensitivity and faster duty cycle of CAD-MS2 could be used for increased phosphopeptide identification by ETD in a nonlinear Paul trap.<sup>17</sup> Alternatively, CAD-MS2 and ETD-MS2 could be used independently to complement each other and potentially reduce the false positive rates provided postacquisition algorithms (PhoS-T-Shunter) are applied.<sup>18</sup> In this work, we focus on validation techniques for identification of pS/pT phosphopeptides from CAD-MS2 spectra acquired in a linear ion trap.

A statistical evaluation of proteomic data is often used to assess the data. While in principle the large-scale analysis of phosphopeptides follows the experimental design of a proteomics assay, there have been few attempts if any to transfer the statistical evaluation of database searches to phosphorylated peptides. Gygi and colleagues reported a large set of 2002 phosphorylation sites.<sup>7</sup> An incipient algorithm that emulates manual validation of phosphopeptides was used to evaluate the post database search data. No statistical analysis of the identifications, however, was presented. Three issues are unique to the identification of differential modifications such as phosphorylation: (1) the dynamics of false positive rates as a function of database size (e.g., number of modified amino acids allowed per peptide), (2) establishing a tandem mass spectrum contains the suspected modification, (3) validating the site identified in a database search.

We develop and compare two methods to validate phosphopeptide spectrum matches. In the first method a  $p$  value is calculated for a match using spectral features of phosphopeptides and statistical multiple testing. The second method uses a support vector machine (SVM) binary classifier, a statistical learning approach, to learn from extracted features of phosphopeptide tandem mass spectra. The trained algorithm is then used to validate a match. SVM has been used to validate SEQUEST search output,<sup>19</sup> to determine charge states for low-resolution tandem mass spectra,<sup>20</sup> and to classify mass spectrometry data of diseased

and normal samples.<sup>21</sup> Neutral loss of phosphoric acid is a common feature of phosphopeptide tandem mass spectra and is often used to substantiate the presence of phosphoserine or phosphothreonine in a peptide sequence. This feature is used predominately to evaluate phosphopeptide identifications in the generation of  $p$  values and for classification by SVM. We show that features associated with phosphate neutral losses have good classifying power. As a working example, we used highly enriched fractions of phosphopeptides from a rat brain nuclear extract. We show that these automatic methods compare well with manual expert validation in a blinded test. The validity of the methods was further confirmed using synthetic peptides containing phosphoserines. We further tested the sensitivity of the SVM-based algorithm for validation of phosphopeptide spectrum matches on phosphopeptides from a HeLa cell nuclear extract.

## EXPERIMENTAL SECTION

**Preparation of Nuclear Extracts from Rat Brain Tissue and HeLa Cells.** Nuclei were pelleted from homogenized brain tissue centrifuged at 1500 rpm for 15 min. Nuclear proteins were extracted with a nuclear/cytosolic fractionation kit (BioVision, Inc.). Protein concentration was determined using a modified Bradford assay (Biorad). A 1 mg sample of nuclear extract from rat brain was further processed by methanol/chloroform extraction. The protein pellet was dissolved by sonication in 50% MeOH in 100 mM Tris, pH 7.6, and digested with trypsin (1:50) at 37 °C overnight. Phosphopeptides were enriched in three fractions by a strategy that will be described elsewhere.

Extraction of phosphopeptides from the HeLa cell nuclear extract followed the same preparation steps.

**Multiple Dimensional Protein Identification Technology (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).

The column was placed in line with an Agilent 1100 quaternary HPLC and analyzed using a 12-step separation. The buffer solutions are 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 and reversal to 100% A in 2 and 3 min of reequilibration 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 to 15% buffer B, and a 97 min gradient from 15 to 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. The final step had the following profile: 3 min of 100% buffer A, 20 min of 100% buffer C, a 10 min gradient from 0 to 15% buffer B, and a 107 min gradient from 15 to 70% buffer B.

Peptides were electrosprayed directly into an LTQ mass spectrometer (ThermoFinnigan, Palo Alto, CA) and analyzed with

(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*; American Society for Mass Spectrometry: Santa Fe, NM, 2006.

(18) Kocher, T.; Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. *J. Proteome Res.* **2006**, *5*, 659–668.

(19) Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. *J. Proteome Res.* **2003**, *2*, 137–146.

(20) Klammer, A. A.; Wu, C. C.; Maccoss, M. J.; Noble, W. S. *Proc. IEEE Comput. Syst. Biotinf. Conf.* **2005**, 175–185.

(21) 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 Biotinf.* **2006**, *7*, 197.

a data-dependent neutral loss routine MS3 (sequential mass spectrometry). A cycle of full scan mass spectrometry (MS) (400–1800  $m/z$ ) was performed followed by MS/MS on the seven most abundant ions. For every  $m/z$  difference of 32.5, 49, and 98 in the three most abundant ions in a tandem MS (MS/MS or MS2) spectrum, a MS3 spectrum was acquired. A normalized collision energy of 35% and an isolation width of 3  $m/z$  units were used for acquisition of both MS2 and MS3 spectra. Each MS2 spectrum was averaged by combining three microscans during acquisition. A total of 100 counts were required as the minimum signal threshold for MS3.

Phosphopeptides from the HeLa cell nuclear extract were analyzed as above except the use of one microscan during acquisition of fragmentation spectra.

**Analysis of MS2 and MS3 Data.** Raw files were extracted with in-house software<sup>22</sup> to produce distinct sets of MS2 and MS3 files. MS2 spectra were filtered using a quality assessment algorithm.<sup>23</sup> Filtered MS2 spectra were searched using SEQUEST<sup>24</sup> with differential modification of +80 Da on STY (phosphorylation), while MS3 spectra were searched with differential modification of –18 Da on ST ( $\beta$ -elimination of phosphate) and M – 48 resulting from the neutral loss of HSOCH<sub>3</sub> (64 Da) from oxidized methionine. We searched the EBI-IPI rat protein database (version 3.05, April 2005) and its reversed sequences. Results from the SEQUEST analysis were statistically evaluated with DTASelect, version 2.0, that incorporates false positive identification rates using decoy matches. MS2 data were analyzed with the following options for DTASelect2.0:<sup>25</sup> –o (removes proteins that are subsets of others); –fp 0.03, selects only peptides with a false positive rate of 3%; –p 1, considers proteins identified by a single peptide; –m 0, displays (for visualization) only the modified peptides; –Smn 10, minimum peptide length of 10 aa; –Smx 25, maximum peptide length of 25 aa. In a separate analysis, MS3 spectra were filtered as above except –fp 0.05, –Smn 7 (minimum peptide length), and –Smx 30 (maximum peptide length).

Data from the HeLa cell nuclear extract were searched by SEQUEST with differential modification of +80 Da on STY with no protease specificity against the EBI-IPI human protein database (version 3.04, March 2005). DTASelect2 filtering criteria allowed for a 5% false positive rate irrespective of the modification status of the peptides.

**Synthetic Peptide Analysis.** Two synthetic peptides, pSFV-LNPTNIGMSKSSQGHVTK (AnaSpec, California) and SFVLNPTNIGMpSKSSQGHVTK (AnaSpec, California), were analyzed by direct infusion to an LTQ mass spectrometer with a 2.5 kV spray potential. Samples were introduced at a constant flow rate of 1  $\mu$ L/min using an infusion pump (Harvard Apparatus). A normalized collision energy of 35% and an isolation width of 3  $m/z$  units were used for acquisition of MS2 spectra.

We also acquired tandem mass spectra for the peptide pSFVLNPTNIGMSKSSQGHVTK at various microscan and injection

time conditions under the above setting. We varied one or three microscans with a 100 or 250 ms injection time to obtain tandem mass spectra of different qualities for the purpose of testing the validation algorithms.

## THEORY

**Multiple Testing and Bonferroni Correction.** We employed multiple testing in the derivation of  $p$  values for peptide phosphorylation validation. Multiple testing is a way of constructing individual confidence levels so that the familywise confidence level can be controlled. With multiple testing, the greater the number of tests performed, the higher the chance that a low  $p$  value will be found for at least the individual tests. The Bonferroni correction is a correction method for performing several dependent or independent statistical tests simultaneously. Explicitly, given  $n$  tests  $T_i$  for hypotheses  $H_i$  ( $1 \leq i \leq n$ ) under the assumption  $H_0$  that all hypotheses  $H_i$  are false, and if the individual test critical values are  $\leq \alpha/n$ , then the familywise critical value is  $\leq \alpha$ . In equation form, if

$$P(T_i \text{ passes} | H_0) \leq \alpha/n \quad (1)$$

for some  $i$ ,  $1 \leq i \leq n$ , then

$$P(T \text{ passes} | H_0) \leq \alpha \quad (2)$$

In our case, the familywise null hypothesis is that the current peptide is not phosphorylated. We then choose the following individual tests. We let A be the event of observing a precursor neutral loss (NL)/base peak (BP) ratio of  $r$ , let B be the event of observing  $n$  number of fragment ion neutral losses, and let C be the event of seeing a  $q$  percentage of unassigned intensity explained by fragment ion neutral losses. We then have individual tests  $T_1$ ,  $T_2$ , and  $T_3$  for events A, B, and C, respectively, where  $P(A) = P(T_1 \text{ passes} | H_0)$ ,  $P(B) = P(T_2 \text{ passes} | H_0)$ , and  $P(C) = P(T_3 \text{ passes} | H_0)$ . The familywise test will then be significant at the  $\alpha$  level if  $P(A)$ ,  $P(B)$ , or  $P(C)$  is less than  $\alpha/3$  by Bonferroni correction or, say, the familywise  $p$  value is 3 times the minimum of  $P(A)$ ,  $P(B)$ , and  $P(C)$ .

To compute the probabilities  $P(A)$ ,  $P(B)$ , and  $P(C)$ , we constructed a large random set: a collection of 5498 spectrum/peptide identifications that are known to be of nonphosphorylated peptides. From the random set we can build empirical random distributions. See the Results and Discussions for details.

**Classification by an SVM.** Here we give a brief introduction to SVM for pattern recognition. Vapnik<sup>26,27</sup> and Cristianini and Shawe-Taylor<sup>28</sup> present an extensive treatment to support vector classification. SVM is one of the supervised statistical learning methods for classification problems, which can also be applied to regression and ranking problems. 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.

(22) 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.

(23) Bern, M.; Goldberg, D.; McDonald, W. H.; Yates, J. R., III. *Bioinformatics* **2004**, *20* (Suppl. 1), I49–I54.

(24) Eng, J.; McCormack, A.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.

(25) Cociorva, D.; Tao, X.; Yates, J. R., III. *54th ASMS Conference on Mass Spectrometry*; American Society for Mass Spectrometry: Santa Fe, NM, 2006.

(26) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Press: New York, 1995.

(27) Vapnik, V. *Statistical Learning Theory*; Wiley Press: New York, 1998.

(28) Cristianini, N.; Shawe-Taylor, J. *An introduction to support vector machines and other kernel-based learning methods*; Cambridge University Press: Cambridge, U.K., 2000.

First, the SVM classifier searches for a hyperplane that separates the two classes with a maximum margin. 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, 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.

Mathematically, the decision function by SVM can be expressed as

$$f(\vec{x}) = \text{sign}\left(\sum_{i=1}^n y_i \alpha_i K(\vec{x}, \vec{x}_i) + b\right) \quad (3)$$

where  $\vec{x}$  represents the input vector and  $y$  its associated output values. Coefficients  $\alpha_i$  can be computed by quadratic programming, while the sign function provides the decision rule.

Our classification study by SVM was implemented using the LIBSVM package.<sup>29</sup> Details of the study are described in the Results and Discussion.

## RESULTS AND DISCUSSION

In this study we develop an approach for identification and validation of phosphopeptides with high confidence. We propose a multitier strategy that assesses the statistical nature of each process along the identification pipeline. False positive rates are estimated by employing a decoy database containing reversed protein sequences.<sup>30</sup> Matches to a tandem mass spectrum are challenged by comparing those derived from two different search conditions. Modification search space is artificially increased by allowing more missed proteolytic cleavages and more modifications per peptide sequence (e.g., three versus six).<sup>31</sup> By comparing the matches between one search state (e.g., three) versus another search condition (e.g., six), only the highest quality and most confident matches are found in both search conditions. To develop a phosphopeptide data set for the development of methods of spectral evaluation, we used the above methods to create a stringently filtered data set. This data set was used to identify spectral features of phosphopeptides for use in validating phosphopeptide/spectrum matches.

**Acquisition of Phosphopeptide Spectra by MudPIT.** Nuclear extracts were prepared from lysates of brain tissue. For further processing, samples were extracted with methanol/chloroform and then digested with trypsin in 50% organic solvent without any reduction/alkylation. To simplify the peptide mixture, we enriched for phosphopeptides through a simple fractionation procedure that will be described elsewhere. Three fractions with a high content of phosphopeptides were analyzed by our standard MudPIT

platform<sup>32</sup> as described in the Experimental Section. We acquired MS2 and on-the-fly neutral loss triggered MS3 spectra for thorough characterization of phosphopeptides, an instrument routine first described by Beausoleil et al.<sup>7</sup>

**Phosphopeptide Data Sets.** This section describes generation of data sets for both classification by SVM and statistical multiple testing.

For each of the three mass spectrometry data sets generated from the rat brain sample, two separated tryptic SEQUEST searches (against the IPI rat protein database, version 3.05, complemented with its decoy) were carried out, each allowing differential phosphorylation modification (mass shift +80) on S, T, and Y, but one allowing up to three modifications per peptide and one missed cleavage site per peptide ("3-mod") and the other allowing up to six modifications per peptide and four missed cleavage sites per peptide ("6-mod"). Tandem mass spectra of identified phosphopeptides in both 3-mod and 6-mod intersection sets were collected with a 3% false positive rate as measured using a decoy database. Phosphopeptides were limited to lengths between 10 and 25 amino acids. A false positive rate was calculated without separation of modified and unmodified peptides; therefore, the same criteria were applied *independent* of spectral appearance. All spectra of phosphopeptides were passed to the next step.

A second SEQUEST search was performed using this smaller set of tandem mass spectra against the original IPI rat protein database under no-enzyme search conditions. Phosphopeptides containing phosphotyrosine residues were discarded. Phosphopeptides identified from this second search that passed the DTASelect filtering criteria constituted the positive sets. We partitioned positive sets into a "positive training set" (376 spectrum/peptide matches) and a "positive testing set" (308 spectrum/peptide matches). Nonphosphorylated peptides (from 3-mod SEQUEST search results) formed the negative sets that were further partitioned into a "negative training set" (376 spectrum/peptide matches) and a "negative testing set" (1011 spectrum/peptide matches).

Phosphopeptide identifications present in only one search condition (e.g., 3-mod) but not in the other (e.g., 6-mod) were collected in an "exclusion set" (684 spectrum/peptide matches). Thus, the phosphopeptide identifications in the exclusion set are not as reliable as the phosphopeptide identifications in the positive set. The exclusion set serves as a useful testing set as shown below.

One set of 5498 nonphosphorylated peptide identifications are also collected from the 3-mod search results. This set serves as the nonphosphorylation random set (as outlined in the Theory) for the purpose of generating a random distribution for several critical features: the precursor neutral loss/base peak ratio, the fragment ion neutral loss count, and the fraction of fragment ion neutral loss intensities of total unassigned peaks.

A total of 281 tandem mass spectra were collected from the direct injection of the synthetic peptides into the LTQ mass spectrometer. The spectra were searched using SEQUEST against a yeast fasta database with the sequence of the synthetic peptide appended. Peptide identifications were then filtered by using the DTASelect program to produce the set of 143 MS/MS spectra corresponding to two different synthetic peptides.

(29) Chang, C. C.; Lin, C. J. LIBSVM: a library for support vector machines. National Taiwan University, 2001; software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

(30) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2*, 43–50.

(31) Moore, R. E.; Young, M. K.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 378–86.

(32) Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R., III. *Nat. Biotechnol.* **2003**, *21*, 532–8.

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

		PNL/BP ratio	mean <i>I</i> of b ions (%)	mean <i>I</i> of y ions (%)	mean <i>I</i> of U peaks (%)	no. of FNLs	mean <i>I</i> of FNLs (%)	percentage of peak counts, FNLs/U (%)	percentage of <i>I</i> , FNLs/U (%)
positive	mean	0.707	3.54	9.57	10.79	6.173	10.49	18.03	18.52
	std dev	0.363	3.18	5.96	6.16	3.159	7.40	8.77	11.73
negative	mean	0.021	8.56	19.97	14.82	1.386	8.91	5.13	6.07
	std dev	0.066	8.12	10.67	7.66	1.758	12.28	5.78	9.45
FCS		1.095	0.022	0.065	0.012	4.661	0.001	0.114	0.073

<sup>a</sup> (1) Precursor neutral loss (PNL)/base peak (BP) ratio; (2) average intensity of b ions; (3) average intensity (*I*) of y ions; (4) average intensity of the 2*L* unassigned (U) peaks; (5) number of fragment ion neutral losses (FNLs); (6) average intensity of fragment ion neutral losses; (7) percentage of unassigned peak intensity explained by fragment ion neutral losses; (8) percentage of unassigned peak counts explained by fragment ion neutral losses. FCS = Fisher criterion score.

**Characterizing Training Data Sets.** For each spectrum/peptide match, we extracted an 8-dimensional feature vector,  $\vec{x} = (f_1, f_2, \dots, f_8)$  in  $R^8$ , where  $f_i$  is the *i*th feature: (1) precursor “NL”/“BP” ratio; (2) average b ion intensity; (3) average y ion intensity; (4) average intensity of the 2*L* unassigned peaks, where *L* is the length of the peptide; (5) number of fragment ion neutral losses; (6) average intensity of fragment ion neutral losses; (7) percentage of unassigned peak intensity explained by fragment ion neutral losses; (8) percentage of unassigned peak counts explained by fragment ion neutral losses.

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 further analyzed, and some features that differ significantly between the positive set and negative set are selected for discussion below.

One prominent feature of tandem mass spectra of phosphopeptides is the dominant intensity of the precursor neutral loss peak (a well-known feature). Most often the precursor neutral loss peak is also the base peak. We studied the distribution of NL/BP ratios between the positive and negative set. Figure 1A shows that, for the positive set, the NL/BP ratios are generally higher, 51% (191 out of 376) of the spectra of which have an NL/BP ratio value of 1 (meaning the precursor neutral loss peak is also the base peak).

The second feature that differentiates well the positive from the negative set is the number of fragment ion neutral losses (also a well-known feature of phosphopeptides). For the positive set, the mean number of fragment ion neutral losses is 6.173, with a standard deviation of 3.159. For the negative set, the mean is 1.386 with a standard deviation of 1.758. The distributions of this second feature for the positive set and negative set are shown in Figure 1B. Neutral losses from fragment ions represent a critical feature of phosphopeptide/spectrum matches. Unambiguous localization of a phosphorylation site is critical for any phosphorylation mapping experiment. In principle, phosphosite localization could be obtained to some confidence level by sequencing b and y ions for singly phosphorylated peptides. However, b/y ions pairs have limited use in localizing phosphosites from multiply phosphorylated peptides displaying clusters of neutral losses.

The third feature that distinguishes the positive from the negative set is the percentage of unassigned peak intensities that could be explained by neutral losses from fragment ions. We annotated each tandem mass spectrum for b ions and y ions, as well as the precursor neutral loss ion. The top 2*L* most intense

unannotated peaks from the experimental spectrum are then classified as unassigned peaks. For each spectrum, we determined how many of these 2*L* unassigned peaks can be explained by fragment ion neutral losses and computed the fragment ion neutral loss intensities over the total intensity for the 2*L* peaks. The distributions for the positive and the negative sets are shown in Figure 1C.

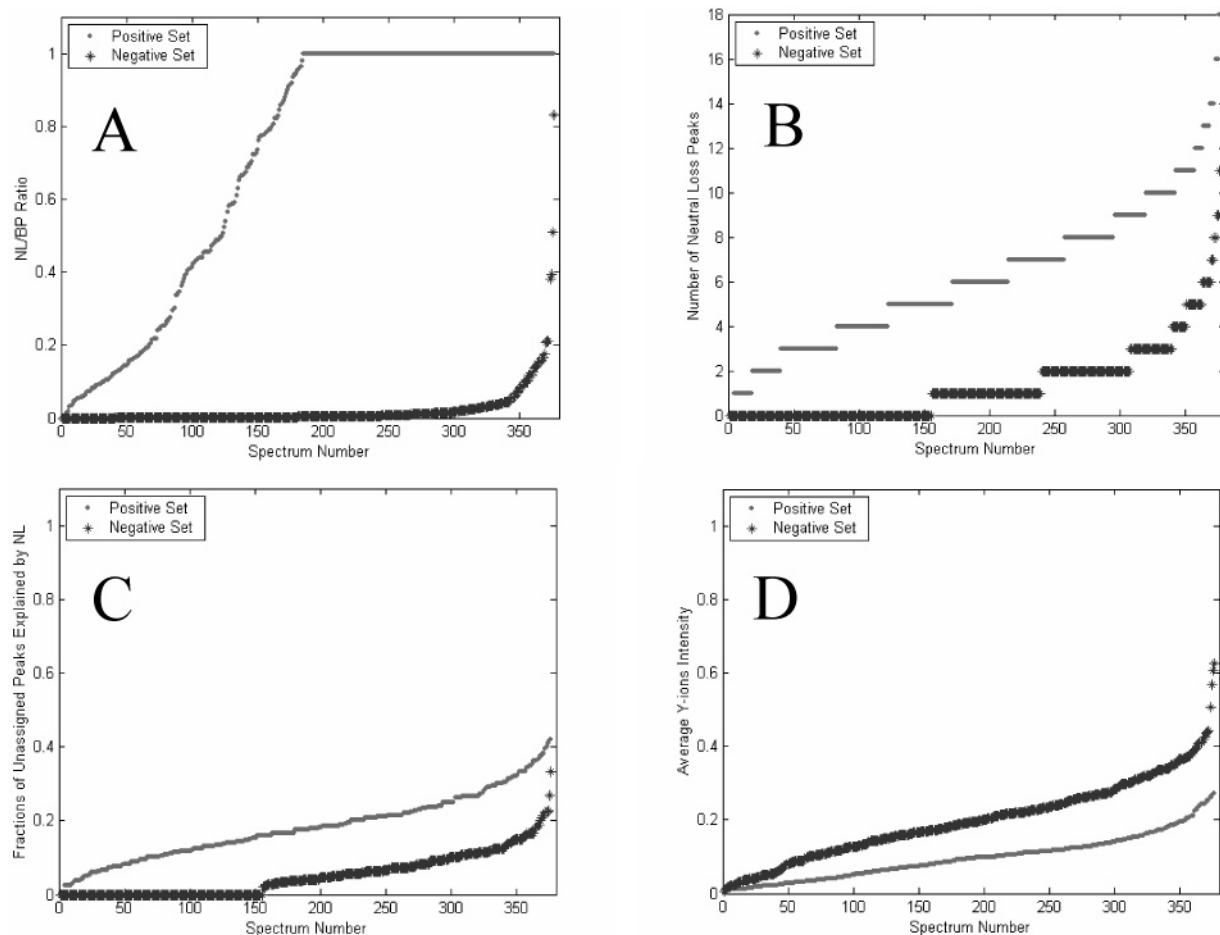
The average y ion intensity, the fourth feature, is quite different between the positive and the negative sets as shown in Figure 1D. This difference could be a result of the dominant intensity of the precursor ion neutral loss peak. We also found that the distribution of the average b ion intensity is different between the positive set and negative set, although the difference is not as prominent (see Supporting Information Figure 1B).

The histograms of all eight extracted features for the positive training set and negative training set are detailed in Supporting Information Figure 1.

**Classifications by SVM.** We used a binary classifier, SVM, to assess the correctness of phosphopeptide identification. Each spectrum/phosphopeptide pair can be classified as one of two classes: correct phosphopeptide identification or incorrect phosphopeptide identification.

We evaluated the classification power of SVM using the eight features described above. We used the LIBSVM package<sup>29</sup> and trained the program using 376 positive (phosphorylated) spectrum/peptide identifications and 376 negative (nonphosphorylated) spectrum/peptide identifications. The following kernel functions were tried: linear kernel function, polynomial kernel function with  $d = 2$ , and radial basis function (RBF). The LIBSVM-based program outputs predictive posterior probability as a predictive value. A predictive value larger than 0.5 means the phosphorylation is positive. A predictive value smaller than 0.5 means the phosphorylation is negative. A predictive value close to 1 means a strong phosphorylation prediction, while a predictive value larger than 0.5 and close to 0.5 means a weak phosphorylation.

The prediction accuracies on the testing sets are shown as ROC (receiver operating characteristic) curves in Figure 2 (also see Supporting Information Table 1). For SVM with the RBF kernel, 95.78% of identifications from the positive testing set are classified as correct phosphopeptide identifications, 97.82% of identifications from the negative testing set are classified as correct unphosphorylated peptides, and 27.78% of identifications from the exclusion set are classified as correct phosphopeptide identifications. The percentage of correct phosphopeptide identifications (as determined by the SVM classifier) in the exclusion set is lower



**Figure 1.** Distribution of extracted features for the training set. The plots show the distributions of features for 944 positive spectrum/peptide identifications and 944 negative spectrum/peptide identifications (randomly selected from the 1064 negative training spectrum/peptide matches). Key: (A) distribution of precursor neutral loss/base peak ratios; (B) distribution of the number of fragment ion neutral losses; (C) distribution of the percentage of unassigned peak intensities that could be explained by fragment ion neutral losses; (D) distribution of the average y ion intensity.

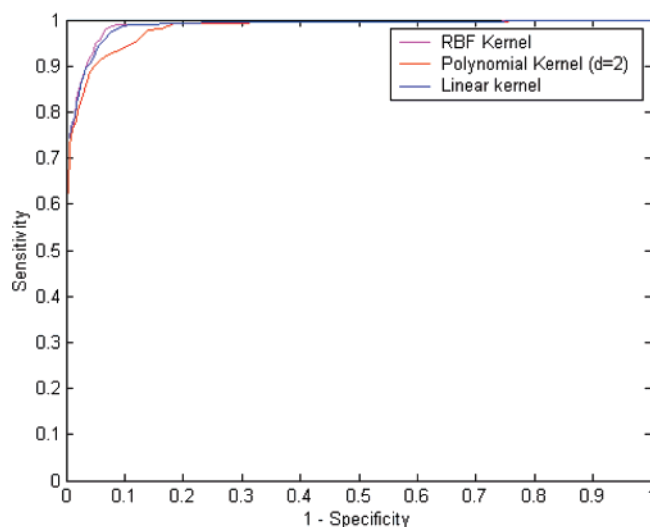
than that of the positive set but higher than that of the negative set.

These findings are consistent with the fact that phosphopeptide identifications in the exclusion set are not as reliable as those in the positive set, as previously discussed. However, there still may be some correct phosphopeptide identifications in the exclusion set, and a correspondingly higher percentage of phosphopeptide identifications are present as compared to those of the negative set. Twenty tandem mass spectra from the positive set and 20 of each of the three categories (strong positive phosphorylation, weak positive phosphorylation, and negative phosphorylation) of the exclusion data set are provided as Supporting Information.

Classifying the testing set using different kernel functions gives different correct phosphopeptide identification percentages (Supporting Information Table 1). In our experiments, SVM classification with the RBF kernel provided the most stringent correct phosphopeptide identifications.

The error rates of SVM classification were estimated by leave-one-out cross-validation. The estimated cross-validation error rates for linear, polynomial ( $d = 2$ ), and RBF kernel functions are 6.52%, 6.25%, and 5.32%, respectively.

One attractive property of SVM is that SVM has the capability to use selected training data points (support vectors) for clas-



**Figure 2.** ROC plots for SVMs using different kernels on the test data sets. The true positive and false positive rates are calculated from the positive test set and negative test set, respectively. The plots show results from the linear kernel SVM, polynomial ( $d = 2$ ) kernel SVM, and RBF kernel SVM.

sification purposes. It is believed that all the information in the training data can be represented by these selected support vectors.

In our study, about 40–45% of the training vectors ended up as support vectors. For example, for the SVM with a second-degree polynomial kernel function, 155 (41%) of the training vectors were selected as support vectors.

To see which features provide the most discriminatory power, the Fisher criterion score (FCS) for each feature was also 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

$$\text{FCS} = \frac{(\mu_A - \mu_B)^2}{\sigma_A + \sigma_B} \quad (4)$$

The higher the FCS score for a feature, the greater the difference between the positive set and the negative set for that feature. A higher FCS score also means that the corresponding feature can provide more discriminatory power for SVM. For example, according to Table 1, the feature that differentiates the positive training set the most from the negative training set is the number of fragment neutral loss ions, with an FCS score of 4.661. The second most predictive feature is the precursor neutral loss to base peak ratio, with an FCS score of 1.095. We tested the performance of the SVM trained with the top five features ranked by FCS scores. The classification power remained almost the same (data not shown), indicating that the top features indeed provide the most important information for the SVM.

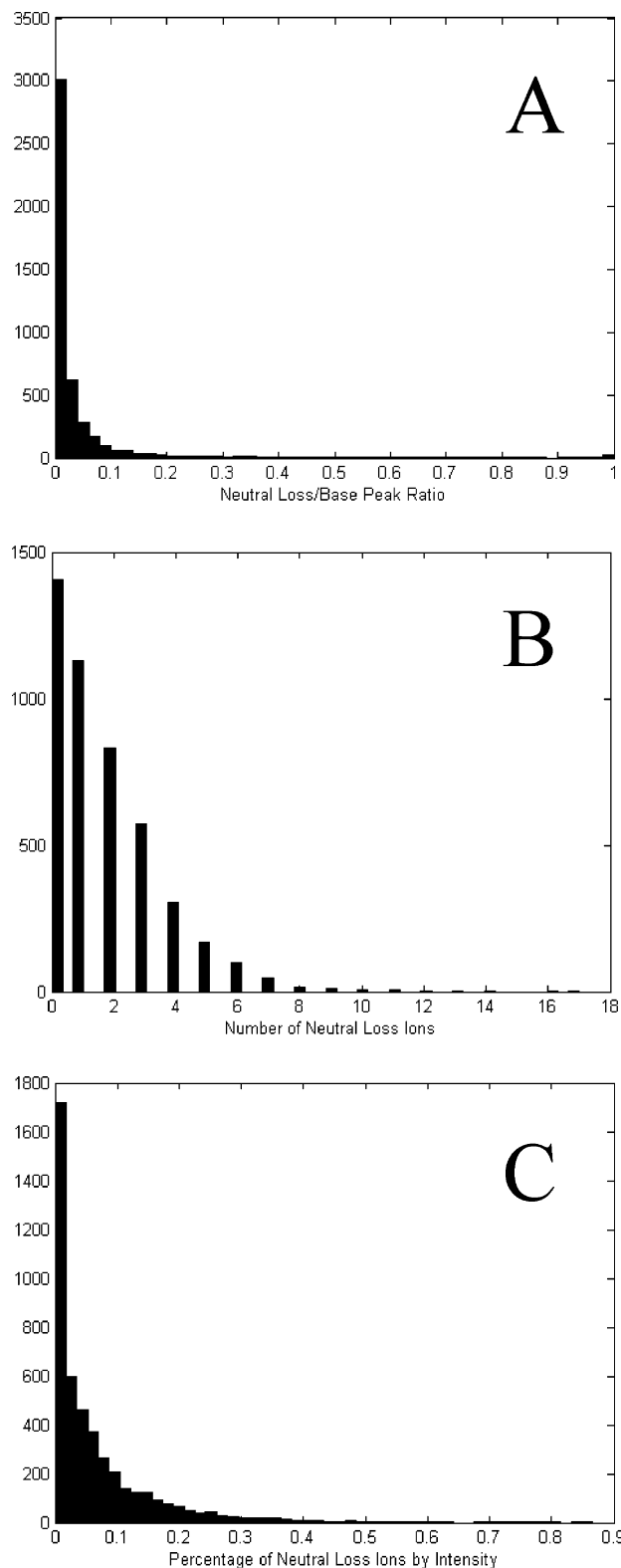
#### Empirical Random Distributions and $p$ Value Calculation.

To compute the  $p$  value for each single event  $P(A)$ ,  $P(B)$ , or  $P(C)$  (as discussed in the Theory), a random distribution is required. We generated the random distributions by using a large collection of nonphosphorylated spectrum/peptide identifications. Consistent with the assumption in our theory, we postulate that the current peptide was not phosphorylated. Parts A–C of Figure 3 show the empirical random distribution for the following features: the precursor neutral loss/base peak ratio, the fragment ion neutral loss count, and the fraction of fragment ion neutral loss intensities for the total unassigned peaks, respectively.

From these empirical random distributions, we can then compute the  $p$  value for each single event. For example, let A be the event of observing a precursor neutral loss/base peak ratio larger or equal to 0.6, then  $P(A)$  = area of bars in the distribution (Figure 3A) with NL/BP larger or equal to 0.6 = 0.0111. After the  $p$  values for each individual single event are calculated, the familywise  $p$  value will then be assessed according to a Bonferroni correction (see the Theory).

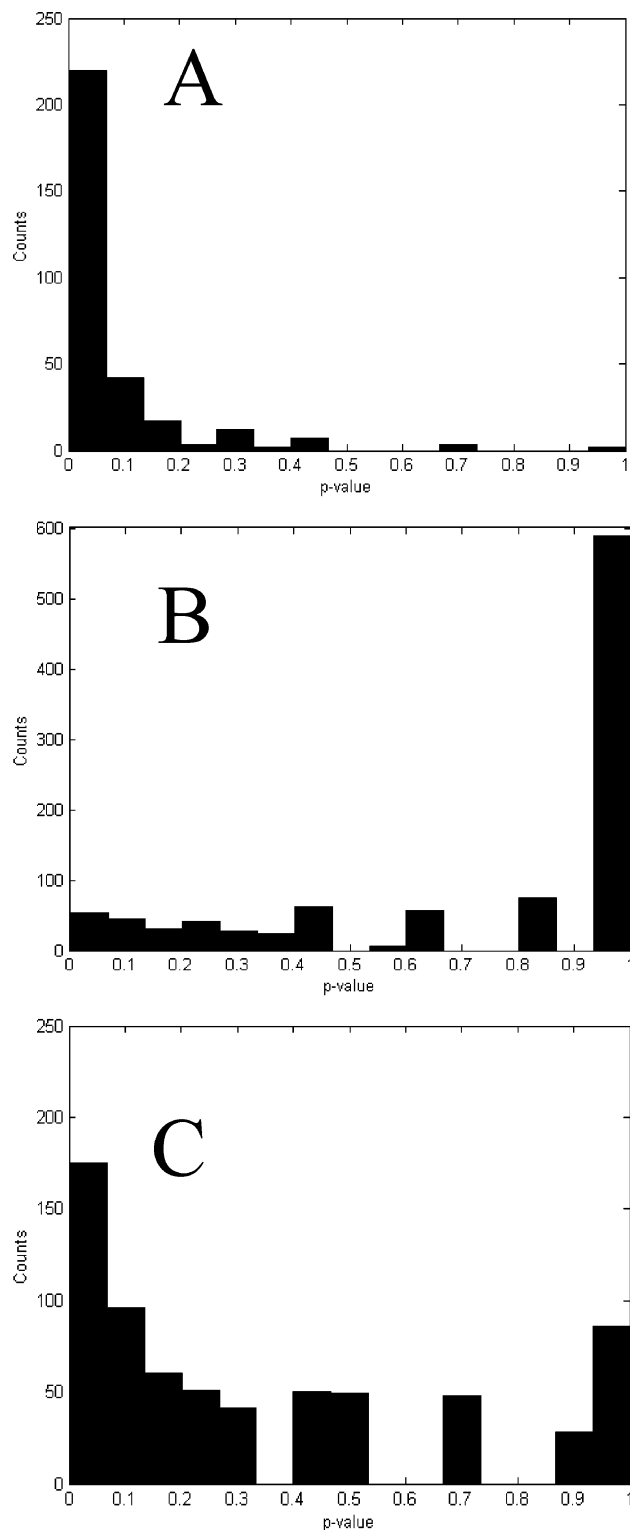
We computed the  $p$  values for the three sets: positive testing set, negative testing set, and exclusion set. The distributions for these three sets are shown in Figure 4. For the positive testing set, we see that most  $p$  values are close to 0 (Figure 4A), confirming the positive set contains a large portion of phosphorylated peptides. For the negative testing set, most  $p$  values are close to 1 (Figure 4B), which is what we expected. For the exclusion set (Figure 4C), approximately 28% of the spectrum/peptide identifications have  $p$  values close to 0. As concluded from Figure 4C, a good portion of phosphopeptide identifications are still present in the exclusion set.

After calculation of the  $p$  values, a  $p$  value cutoff can be set to screen for correct identifications. Table 2 shows the results on



**Figure 3.** Empirical random distributions for extracted features (sample size 5498): (A) random distribution for precursor neutral loss/base peak ratios; (B) random distribution of the number of fragment ion neutral losses; (C) random distribution of the percentage of unassigned peak intensities that could be explained by fragment ion neutral losses.

the testing sets using two  $p$  value cutoffs: 0.05 and 0.01. At the  $p$  value 0.05 cutoff, 82.79% (255/308) of the spectrum/peptide identifications from the positive testing set are classified as



**Figure 4.** Distribution of  $p$  values for the testing sets: (A) distribution of  $p$  values calculated from the positive testing set; (B) distribution of  $p$  values from the negative testing set; (C) distribution of  $p$  values from the exclusion set.

phosphorylated peptides, with 91.10% (921/1011) of the spectrum/peptide identifications from the negative testing set classified as unphosphorylated peptides. When the  $p$  value cutoff is lowered to 0.01, although only 68.50% (211/308) of the spectrum/peptide identifications from the positive testing set are classified as phosphorylated peptides, 97.73% (988/1011) of the spectrum/

**Table 2. Correctness (%) of Assignments by Multiple Testing**

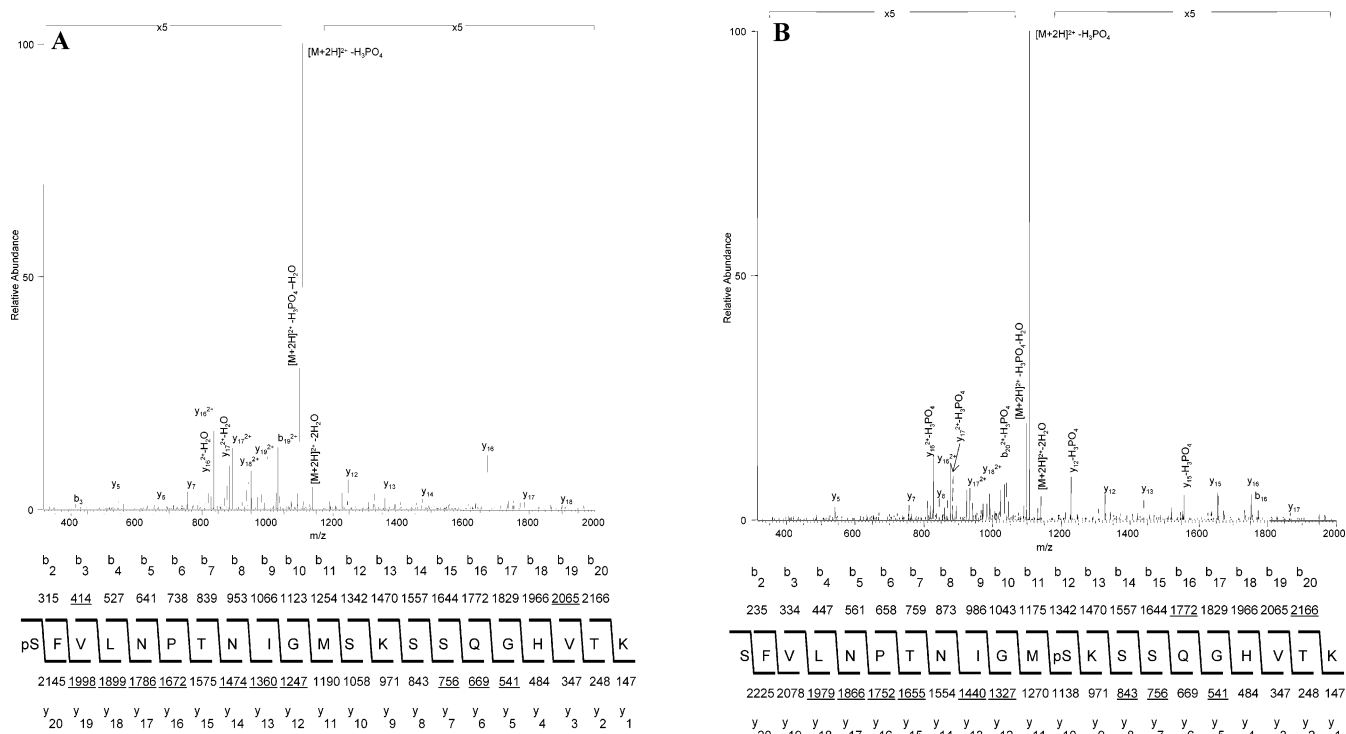
$p$ value cutoff	positive set	negative set	exclusion set
0.05	82.79	8.92	23.25
0.01	69.48	2.27	9.94

peptide identifications from the negative testing set are classified as unphosphorylated peptides. The phosphorylated peptide portions for the exclusion set are 23.25% (159/684) and 9.94% (68/684) for  $p$  value cutoffs 0.05 and 0.01, respectively, again indicating that the quality of phosphopeptide identifications for this set contains features from both the positive set and the negative set.

**Comparing the Multiple Testing Method to SVM.** We compared the multiple testing method to the SVM method on the positive testing set. By using the 0.05  $p$  value cutoff for the multiple testing method and the RBF kernel for the SVM method, we found that the two methods agree with each other 85% of the time (262 out of 308 spectrum/peptide identifications, Supporting Information Figure 2). The SVM method produces more true positive and less false positive results in general (Table 2 and Supporting Information Table 1). For those spectra where the two methods do not agree with each other, the  $p$  values and the SVM predictive values are in the marginal area. We find the SVM method is more sensitive and more specific in this study.

**Validation of the Methods with a Blinded Test.** We also carried out a blinded test to validate the algorithms. Twenty-four spectrum/peptide identifications from the exclusion set were selected, and the  $p$  values and class labels (correct or incorrect phosphorylation identifications) were masked. For these 24 phosphopeptide identifications, the multiple testing method (with the  $p$  value 0.05 cutoff) and the SVM method (with the RBF kernel) agree with each other very well, except for 2 spectrum/peptide identifications where the two methods do not agree with each other. Again both sets of scores are in the marginal area of  $p$  values and SVM predictive values, as shown in Supporting Information Figure 3. These masked spectrum/peptide identifications were then presented to a human expert for manual validation of peptide phosphorylation. Supporting Information Figure 3 shows the manual validation results. There are three spectrum/peptide identifications where manual validation does not agree with the SVM method. The same number of disagreements is also observed when the manual validation method is compared to the multiple testing method. In each case, two of the three disagreements are with spectrum/peptide identifications of marginal scores, while the third disagreement (see Supporting Information Figure 3) has a low  $p$  value and a high SVM predictive value. An MS3 spectrum was also observed for this third identification. A preponderance of evidence suggests that manual validation was incorrect and the automated methods were correct on this particular identification.

**Validation of the Methods with Synthetic Peptides.** We further validated the two methods using the synthetic phosphopeptides pSFVLNPTNIGMSKSSQGHVTK and SFVLNPTNIGMpSKSSQGHVTK. These two peptides have the same amino acid sequence. Both contain one phosphoserine, but at different positions within the peptide, as indicated by the lower case "p".



**Figure 5.** Example spectra from synthetic phosphopeptides: (A) example spectrum for synthetic peptide pSFVLNPTNIGMSKSSQGHVTK; (B) example spectrum for synthetic peptide SFVLNPTNIGMpSKSSQGHVTK.

**Table 3. XCorr, SVM Predictive Posterior Probability Value, and Multiple Testing *p* Values for Phosphopeptide Spectrum Matches under Different Acquisition Conditions**

	XCorr		SVM (RBF) probability		multiple testing <i>p</i> value	
	mean	std dev	mean	std dev	mean	std dev
one microscan, 100 ms injection time	1.9107	0.3946	1	3.77E-04	0.0648	0.0794
one microscan, 250 ms injection time	2.3358	0.4511	0.996	1.29E-02	0.0715	0.0971
three microscan, 100 ms injection time	2.7479	0.4207	0.995	1.46E-02	0.0236	0.0275
three microscan, 250 ms injection time	2.9674	0.4556	1	1.79E-04	0.0154	0.0199

A total of 281 tandem mass spectra were searched using SEQUEST against a yeast fasta protein sequence database with the synthetic peptide appended. DTASelect filtering on the SEQUEST search results gave rise to 143 redundant spectrum/peptide identifications corresponding to the two synthetic phosphopeptides.

The correctness of these 143 redundant spectrum/peptide identifications was assessed by the SVM classifier and multiple testing method. For the SVM classification with either one of the three kernel functions (linear, polynomial kernel function of degree  $d = 2$ , or RBF), all identifications were classified as positive. All *p* values computed were smaller than the 0.05 cutoff. The results confirm the validity of the methods. Example spectra for the two synthetic peptides are shown in Figure 5.

**Sensitivity and Specificity of the Multiple Testing and SVM Methods.** To illustrate the sensitivity of the multiple testing and SVM methods, we performed the following two experiments.

In experiment one, we varied microscan and injection time conditions in an infusion experiment for the synthetic peptide pSFVLNPTNIGMSKSSQGHVTK to obtain tandem mass spectra of different qualities.<sup>33</sup> The results are shown in Table 3. The mean

and standard deviation for SEQUEST XCorr scores for tandem mass spectra acquired under one microscan and a 100 ms injection time are 1.9107 and 0.3946, respectively. The mean XCorr score is increased to 2.9674 when we set the acquisition conditions to three microscans and increase the injection time to 250 ms. We obtained intermediate mean XCorr scores of 2.3358 and 2.7479, using one microscan and a 250 ms injection time condition and three microscans and a 100 ms injection time condition, respectively. The SVM algorithm gives consistently high predictive posterior probability scores. Multiple testing also gives consistently low *p* values for this synthetic peptide data set. Overall, both methods are more sensitive in terms of detecting phosphopeptide spectrum matches (also see Supporting Information Table 2).

In experiment two, we tested the classification power of the SVM with the RBF kernel on a large data set acquired from a HeLa cell nuclear extract. After SEQUEST analysis and DTASelect filtering, we obtained 6332 forward phosphopeptide/spectrum identifications with 313 reverse phosphopeptide/spectrum identifications, i.e., a forward false positive rate of 4.94% (Table 4). Further processing of this data set by SVM (RBF kernel) reduced the false positive rate by 50% while retaining more than 90% of

(33) Venable, J. D.; Yates, J. R., III. *Anal. Chem.* **2004**, *76*, 2928–37.

**Table 4. Reducing the False Positive (FP) Rate by SVM (with the RBF Kernel) after DTASelect**

DTASelect			SVM (with RBF kernel)		
forward	reverse	FP rate (%)	forward	reverse	FP rate (%)
6332	313	4.94	5740	124	2.16

the forward hits. Consequently, SVM (RBF kernel) showed increased power for post-SEQUEST filtering of phosphopeptides.

**Test of the SVM Method on OMSSA Search Results.** The features we used for the validation of phosphopeptide identification are not derived from SEQUEST scores. Thus, the methods we developed here should be search engine independent and will be widely applicable to proteomics studies. To prove this, we tested the use of the SVM method (with the RBF kernel) on OMSSA<sup>34</sup> search results. The 308 MS/MS spectra from the positive testing set were used to search against the IPI rat protein database, version 3.05, using OMSSA and allowing differential phosphorylations on S, T, and Y. OMSSA returned 219 phosphopeptide identifications. The SVM classifier trained using SEQUEST results was then applied. Out of these 219 phosphopeptide identifications, 212 (or 96.80%) were classified by the SVM classifier as correct phosphopeptide identifications. This shows that our method could be widely applicable to proteomic studies of phosphorylation using data from linear ion trap mass spectrometers.

## CONCLUSIONS

We developed two methods for the automatic validation of phosphopeptide identifications using a linear ion trap mass spectrometer and SEQUEST database searches. By using statisti-

(34) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. *J. Proteome Res.* **2004**, *3*, 958–964.

(35) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. *Nat. Biotechnol.* **2006**, *24*, 1285–92.

(36) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. *Cell* **2006**, *127*, 635–648.

cal multiple testing, we can calculate the  $p$  value for each tentative peptide phosphorylation. In a second method, we use a binary classifier to predict whether a tentative peptide phosphorylation is true. Both methods were developed for the validation of phosphoserine- and/or phosphothreonine-containing peptide identifications with one or two phosphorylations per peptide. We used highly enriched fractions of phosphopeptides from a rat brain nuclear extract. We show that these two methods agree well with each other with better performance by the SVM method. We also show that these automatic methods match very well with manual expert validation in a blinded test. Additionally, the algorithms were tested on the identification of synthetic phosphopeptides.

We focused on the validation of phosphorylated peptides that were identified by the SEQUEST search engine, but demonstrated the approach is applicable to other search algorithms. Since some phosphorylated peptides might have low fragmentation efficiency over the peptide chain backbone, the identification of such peptides will still be a challenge. The SVM and statistical multiple testing approaches developed simply assess whether a search result is likely to be phosphorylated. We are also examining strategies to assess the predicted sites of modification.

Note: While this paper was under review, two papers addressed the statistical evaluation of identifications of phosphopeptides.<sup>35,36</sup>

## ACKNOWLEDGMENT

B.L. and C.R. contributed equally to this paper. C.R. and J.Y. acknowledge support from NIH Grant 5R01MH067880-02 and NIH Grant P41 RR11823-10. B.L. is supported by CFFT computational fellowship BALCH05X5. We thank Dr. Daniel Cociorva and Dr. Greg Cantin for helpful reading of the manuscript.

## SUPPORTING INFORMATION AVAILABLE

Various experimental data, figures, and tables. This material is available free of charge via the Internet at <http://pubs.acs.org>.

Received for review July 21, 2006. Accepted December 7, 2006.

AC061334V