Published in final edited form as:

J Proteome Res. 2008 July ; 7(7): 2803–2811. doi:10.1021/pr700773p.

PhosphoScan: A Probability-Based Method for Phosphorylation Site Prediction Using MS2/MS3 Pair Information

Yunhu Wan^{*,†}, Diane Cripps[†], Stefani Thomas^{†,‡}, Patricia Campbell[§], Nicholas Ambulos^{†,§}, Ting Chen^{||}, and Austin Yang^{†,⊥}

[†]Greenebaum Cancer Center, University of Maryland, Baltimore, Maryland 21201

[‡]Department of Radiation Oncology, University of Maryland, Baltimore, Maryland 21201

[§]Department of Microbiology and Immunology, University of Maryland, Baltimore, Maryland 21201

^{||}Department of Computational Biology, University of Southern California, Los Angeles, California 90089

[⊥]Department of Anatomy and Neurobiology, University of Maryland, Baltimore, Maryland 21201

Abstract

Phosphopeptide identification and phosphorylation site localization are crucial aspects of many biological studies. Furthermore, multiple phosphorylations of peptides make site localization even more difficult. We developed a probability-based method to unambiguously determine phosphorylation sites within phosphopeptides using MS2/3 pair information. A comparison test was performed with SEQUEST and MASCOT predictions using a spectral data set from a synthetic doubly phosphorylated peptide, and the results showed that PhosphoScan analysis yielded a 63% phosphopeptide localization improvement compared with SEQUEST and a 57% improvement compared with MASCOT.

Keywords

phosphorylation; PTM (post translational modification); protein identification; MS2; MS3; NL (neutral loss)

Introduction

Protease digestion followed by LC–MS/MS analysis has conventionally been utilized to identify the sequences of proteins obtained from biological samples. A number of programs are available that utilize the resulting spectra for both *de novo* and database-based protein sequencing; SEQUEST¹ and MASCOT² are two such popular database-search programs for identification of peptide and protein sequences. Now that these methodologies for identification of amino acid sequences are well established, scientific interest is shifting

*To whom correspondence should be addressed. ywan@som.umaryland.edu.

toward the identification of protein post-translational modifications (PTMs), giving rise to a need for computational tools that complement existing programs.^{3,4}

Among the wide variety of post-translational modifications, protein phosphorylation is one of the most critical regulators of biological function. It has been estimated that, within the eukaryotic cell, approximately 30% of proteins are phosphorylated at any given point.⁵ Phosphorylation indispensably governs the cellular processes of differentiation, protein localization, and protein–protein interactions. Large-scale phosphorylation analyses should thus prove to be highly useful in identifying pathways involved in cellular processes.

Phosphorylation can occur on serine, threonine, or tyrosine residues. Collision-induced dissociation (CID) of phosphopeptides, in addition to providing spectra that can be used for sequence identification, results in the loss of phosphoric acid (H_3PO_4 , 98 Da) from phosphoserine and phosphothreonine. This neutral loss can be used as a marker for the presence of phosphopeptides. Moreover, an additional spectrum that provides complementary sequence information can be obtained by triggering a data-dependent MS3 scan using the neutral loss peak resulting from the loss of phosphoric acid from the MS2 precursor ion. These MS3 spectra are often of better quality than the MS2 spectra of phosphopeptides.

Several phosphopeptide and phosphoprotein databases have been compiled. Currently, Phospho.ELM (version 6, phospho.elm.eu.org)⁶ lists 13 613 experimentally verified phosphorylation sites for 3674 different proteins; Phospho-Site (www.phosphosite.org)⁷ lists 6084 nonredundant phosphorylation sites on 2430 proteins, and Phosida (www.phosida.com)⁸ lists 6600 phosphorylation sites on 2244 proteins.

The correct identification of peptides with PTMs involves an increased level of complexity over identification of amino acid sequence alone. The amount of computational time dedicated to PTM analyses depends upon several factors: the total number of spectra within a data set, database size, and the number and types of modifications considered. The CPU time required for a PTM search with phosphorylation options may reach several times that of a non-PTM search, indicating that use of PTM options may increase search time exponentially. By contrast, search time varies only linearly with the number of spectra. Other factors such as the number of missed cleavages and protein mass differences minimally impact the search time. Meanwhile, altering the enzyme digestion parameters has an impact equivalent to changing the database size. A nonspecific digestion option can significantly lengthen the search time (to levels beyond those required for a PTM search).

How to correctly localize phosphorylation sites (phosphosites) on phosphopeptides found in biological samples, particularly on doubly phosphorylated peptides, remains a significant challenge. To this end, Beausoleil et al.⁹ present a method utilizing a probability assignment termed Ascore in order to correctly localize phosphorylation sites. Lu et al.¹⁰ have reported a multiple testing method, coupled with an SVM method that validates phosphorylation prediction results based upon SEQUEST search results. These methods show great promise for improving the localization of phosphorylation sites in proteins from complex biological samples. An additional factor that could also be considered in phosphorylation site

localization studies is MS2/MS3 information. Such information is obtained when neutral loss MS3 scans are performed and is dependent upon detection in the MS2 spectrum of a neutral loss indicating the presence of a phosphopeptide; the combined information from the MS2 and MS3 spectra might possibly be utilized to ascertain the correct location(s) of phosphorylation sites. To this end, Zhang et al.¹¹ generated a strategy to intersect MS2 and MS3 spectra in order to run *de novo* sequencing-based algorithms. As an alternative approach, recently, Ulintz et al.¹² developed a probability based method to combine MS2 and MS3 information; however, their method is focused on phosphopeptide identification rather than phosphorylation site localization. At the same time, their methods only selected top peptide match for each spectrum. It is quite possible that MS2 and MS3 phosphopeptide identifications are not in the top list.

In this paper, we provide a probability-based method termed PhosphoScan, which is based on MS2/3 pair information for both detecting phosphorylated peptides and pinpointing the exact location(s) of their phosphorylation sites. Using PhosphoScan to analyze data generated from the mass spectrometric analysis of a known hyperphosphorylated protein and a synthetic phosphopeptide, we were able to identify more phosphopeptides with confident phosphorylation site localizations as compared to using conventional MS2 only based approaches, which also resulted in a large number of erroneous phosphopeptide identifications.

Materials and Methods

Affinity-Purification of PHF-Tau and Enzymatic Digestion

Tau protein was extracted and purified from Alzheimer's disease (AD) brain by immunoprecipitation with an MC1 monoclonal antibody, as previously described.¹³ Aliquots of paired-helical filament PHF-Tau, containing approximately 50 μ g of protein, were digested at 37 °C in 40% methanol, 100 mM NH_4HCO_3 at pH 8.5 by three additions of 1% (w/w) trypsin (Promega) or 1% (w/w) Lys-C (Sigma). Digestion was performed by the addition of enzyme for two 2 h periods of incubation, followed by an overnight incubation. The digestion reaction was stopped by the addition of 1% (v/v) acetic acid.

Synthesis of Custom Peptides

The custom peptide STGS-pATSLASpQGER was synthesized on a Prelude peptide synthesizer (PTI Instruments), using Fmoc-derivitized resin (AnaSpec). All amino acids were purchased from AnaSpec. The primary solvent used included a 1:1 ratio of *N,N*-dimethylformamide (DMF) and dichloromethane (DCM). The Fmoc group was removed using 20% piperidine in DMF containing 0.1 M *N*-hydroxybenzotriazole (HOBt). All solvents (Optima grade), piperidine and diisopropylethylamine (DIEA) were purchased from VWR Scientific. HOBt and *O*-(7-Azabenzotriazole-1-yl)-*N,N,N',N'*-tetramethyluronium hexafluorophosphate (HATU) were purchased from Applied Biosystems. Coupling of incoming Fmoc amino acids was achieved using HATU with DIEA as the base in a 1:1:4-fold excess with respect to the resin with 1 h double coupling.

The peptide was cleaved from the resin on the instrument using 95% TFA, 2.5% water, and 2.5% triisopropylsilane for 6 h with nitrogen bubbling at room temperature, then filtered through the reaction vessel into a collection vial. The resin was washed once with cleave solution and then filtered into the same collection vial. The peptide was precipitated by the pouring of pooled filtrates into ice-cold ether. After storage for one hour at -20°C , the precipitated peptide was centrifuged and the supernatant was decanted. The pellet was triturated 3 times with ice-cold ether and air dried. After solubilization in 5% acetonitrile:H₂O, the peptide was filtered, analyzed by HPLC and mass spectrometry, and then lyophilized.

LC-MS/MS and MS/MS/MS Analysis

Affinity-purified PHF-tau peptide digests were separated by a true nanoflow Xtreme-Simple LC system (Micro-Tech Scientific). One run was performed using a $15\text{ cm} \times 75\text{ }\mu\text{m}$ C-18 reverse-phase (RP) column ($5\text{ }\mu\text{m}$ particles, $300\text{ }\text{\AA}$ pore size) (Micro-Tech Scientific) with a linear gradient of 5–25% solvent B (95% acetonitrile, 0.1% formic acid) over 60 min (solvent A was 2% acetonitrile, 0.1% formic acid); another run was performed on a Micro-Tech Scientific $100\text{ cm} \times 75\text{ }\mu\text{m}$ C-18 RP column ($3\text{ }\mu\text{m}$ particles, $300\text{ }\text{\AA}$ pore size) over a gradient of 350 min from 5–25% solvent B. MS2 and MS3 analysis of PTMs were performed on a Thermo Electron LTQ linear ion trap mass spectrometer (Thermo Electron) equipped with a nanospray ion source (spray voltage 2.0 kV, capillary temperature 200°C) using an uncoated $10\text{ }\mu\text{m}$ -ID SilicaTip PicoTip nanospray emitter (New Objective). MS2 and MS3 spectra were acquired using Xcalibur 2.0 SR2 software with the five most abundant ions in each MS scan selected for an MS2 event, subject to dynamic exclusion (exclusion for 30 s after detection twice in 30 s). MS3 scans were triggered if, among the three most abundant ions in the MS2 scan, a neutral loss of 98, 49, or 32.7 Da was detected (corresponding to loss of phosphoric acid on singly, doubly and triply charged precursor ions). Other mass spectrometric data generation parameters were as follows: full scan MS1 mass range = $400\text{--}1800\text{ }m/z$, minimum MS2 signal = 500 counts, minimum MS3 signal = 100 counts, collision energy = 24% for MS2 scans and 35% for MS3 scans, activation time = 120 ms for MS2 scans and 30 ms for MS3 scans.

The custom peptide STGSpATSLASpQGER was analyzed three times as follows. First, 2 pmol of peptide were loaded onto a $15\text{ cm} \times 75\text{ }\mu\text{m}$ C18 ($5\text{ }\mu\text{m}$ particle size, $300\text{ }\text{\AA}$ pore size) column (Micro-Tech Scientific) and eluted over a 30 min gradient of 5–40% solvent B (solvent A: 2% acetonitrile, 0.1% formic acid; solvent B: 95% acetonitrile, 0.1% formic acid) using an XtremeSimple ultra high pressure LC (UPLC) system (MicroTech Scientific) into an LTQ linear ion trap mass spectrometer (Thermo Scientific) utilizing a dynamic nanospray ESI probe (positive polarity, source voltage 1.0 kV, capillary temperature, 200°C). In the instrument method, a full MS scan ($400\text{--}1800\text{ }m/z$) was followed by MS/MS scans of the top 5 ions of the full MS scan (CID parameters: minimum MS signal count = 500, isolation width = $3.0\text{ }m/z$, normalized collision energy = 24%, activation time = 120 ms). An MS/MS/MS scan was triggered upon detection of a NL of $32.66\text{ }m/z$, $49.00\text{ }m/z$ or $98\text{ }m/z$ from among the top 3 product ions of the MS/MS scan (CID parameters: minimum MS/MS signal count = 100, isolation width = $3.0\text{ }m/z$, normalized collision energy = 35%, activation $Q = 0.250$, activation time = 30 ms). Next, a $33\text{ pmol}/\mu\text{L}$ solution of the peptide in

2% formic acid was directly infused into the mass spectrometer and spectra were collected using the same instrument method. Spectra were collected in the same manner upon infusing a 3.3 pmol/ μ L solution of the peptide.

Results

As depicted in Figure 1, three steps are performed to achieve better phosphopeptide predictions using MS2/3 pair information: (1) reduction of spectral data set size by combining similar MS2/MS3 spectra pairs, (2) identification of phosphopeptides by database search, and (3) phosphorylation site localization using a location probability scoring method.

Reduction of Spectral Data Set Size by Combining Similar MS2/MS3 Spectra Pairs

The first step in our analysis of spectral data was the combination of the charge state specific data files that were generated for each MS2 and MS3 spectrum. Because these files differ only with regard to the assumed parent mass, which is derived from the m/z value of the precursor ion and the assumed charge state, the charge 2 and charge 3 data files for each MS2 and MS3 spectrum were counted as one data file.

Next, we identified the true MS2/MS3 spectra pairs. In the mass spectrometry-based analysis of phospho-peptides, MS3 scans are triggered upon the detection of a neutral loss of phosphoric acid during the MS2 fragmentation of a precursor ion. However, in some cases, neutral losses do not result from the loss of phosphoric acid from an actual phosphopeptide. Although reducing a data set to MS2/3 spectra pairs alone decreases the database search space, some MS2/3 pairs remain that are generated randomly from nonphosphopeptides. Thus we employed a method to filter the true MS2/3 pairs from the original data set in order to further reduce the database search space. The resulting phosphopeptide identifications were then based on the filtered MS2/3 pairs.

The components of true MS2/3 spectra pairs should be more similar to each other than to those of a random pair of spectra, given that the MS2 and MS3 spectra of a phosphopeptide are generated from the same original precursor ion. A spectral distance score is needed to prove this hypothesis.

Selecting a good pairwise metric is the key to differentiating real MS2/3 pairs from random pairs. A discriminating method of calculating the pairwise distance will result in decreased distances between similar spectra and increased distances between dissimilar spectra. If a pair of spectra is viewed as two vectors, the distance between them can be defined as the inner product of those two vectors. Therefore, to facilitate the comparison of distinct spectral distances, we have chosen a cosine metric:¹⁴

$$\text{Score} = \frac{\sum_{i=1}^k s_1(i)s_2(i)}{\sqrt{\sum_{i=1}^k s_1^2(i)s_2^2(i)}} \text{ if } |m_1(i) - m_2(i)| < \varepsilon \quad (1)$$

Here $m_1(i)$, $m_2(i)$ represent fragment ions from spectra 1 and 2 with matching m/z values, $s_1(i)$, $s_2(i)$ are their corresponding intensities, k is the number of similar peaks between spectrum 1 and spectrum 2, and ε is a tolerance value set for fragment ion matching. In this analysis, we use $1/m/z$ as the value for ε .

When this method was applied to a data set consisting of 33 930 original spectra originating from analysis of a PHF-Tau spectral data set (see Materials and Methods) filtering of the MS2/3 pairs reduced the data set to 1029 spectra with 513 spectra pairs. The filter threshold we used was chosen based on a distance distribution generated from random MS2/3 pairs and real phosphopeptide MS2/3 pairs. Using the pairwise distance filter described above, 404 spectral pairs were ultimately retrieved from the original 513. The histograms of the distance scores for these spectral pairs (true and random MS2/3 pairs) are presented in Figure 2. An advantage of reducing the total number of spectra from 33 930 to 404 spectra pairs was a dramatic decrease in the total CPU time needed for the database search.

Identification of Phosphopeptides by Database Search

After MS2/3 spectra pairs are selected for processing, peptide identifications are generated for each spectrum. Any database search program may be used here as long as phosphorylation search options are employed. Here, we utilized a previously developed database search method (PepHMM),¹⁵ which identifies peptides by a Hidden Markov Model (HMM) method that defines a score for each sequence assigned to a spectrum by combining machine accuracy, mass peak intensity, and correlation among ions. For this analysis, the database was reindexed and the data set was searched using PepHMM with phosphorylation options of S, T +80 and -18 with no consideration given to the stage of MS fragmentation, e.g., MS2 or MS3. To validate phosphopeptide candidates S, T +80 modifications were considered for MS2 spectra, whereas both S, T +80 and -18 modifications were considered for MS3 spectra. A modification of +80 corresponds to the addition of HPO_3 to a serine or threonine residue, while a -18 modification corresponds to the difference in mass between serine and dehydroalanine or threonine and 2-aminodehydrobutyric acid; dehydroalanine and 2-aminodehydrobutyric acid result from the collision-induced dissociation of phosphoric acid from a phosphoserine or phosphothreonine residue, respectively. Using the score distribution from a control data set, the HMM scores were converted into probability scores (data not shown).

The framework that we used for the phosphopeptide identification is detailed here and is diagrammed in Figure 3. Two types of possible correct MS2/3 combinations were considered. First, if an MS2 spectrum has a +80 single modification on a serine or threonine, the corresponding MS3 spectrum must have a -18 modification on the same serine or threonine. Second, if the MS2 spectrum contains two +80 modifications of serine or threonine, in the corresponding MS3 spectrum one of these sites must undergo a -18 modification while the other must retain the +80 modification. PhosphoScan does not consider more than 2 phosphorylation modifications in a single peptide. Although peptides with more than two phosphorylations are sometimes encountered, this is an uncommon occurrence and the predicted sites of phosphorylation are frequently spurious. Also, PhosphoScan does not consider unphosphorylated peptides. Since we are correlating MS2/3

pairs and the two precursor ions in question have a mass difference of 98, it should not be necessary to consider the case in which the peptide is unmodified. Our method for predicting phosphopeptides is therefore complementary to traditional database search algorithms that focus on phosphopeptide identification.

Phosphorylation Site Localization Using a Location Probability Scoring Method

Full characterization of phosphopeptide sequences includes the identification of both the amino acid sequence and the phosphorylation site(s). However, a limitation of many current database search algorithms is that although the amino acid sequence may be correctly identified, in some cases, the exact phosphorylation site cannot be definitively determined. In addition, scores are retrieved independently for MS2 and MS3 spectra. For instance, on a doubly phosphorylated peptide, SEQUEST may assign similar XCorr values to peptides that have identical amino acid sequences but that have several potential phosphorylation sites, thus resulting in ambiguous phosphorylation site localization. In order to facilitate the unambiguous determination of phosphorylation sites, we developed a localization probability score that is specific to the identification of phosphorylated residues. This probability score takes advantage of the fact that the MS2 and MS3 spectra originate from the same phosphopeptide (Figure 4).

This location probability score is based on the idea that the most probable phosphorylation site will correspond to the residue for which there is a maximum number of pairs of major product ions that correctly bracket that phosphorylation site (such pairs may appear in either the MS2 spectrum or MS3 spectrum or both). Therefore, for each candidate phosphorylation site, ion pairs that are counted are only those for which fragment ion matches exist both before and after the site of phosphorylation. Only one such “ion pair” is counted regardless of how many fragment ions can be used to discriminate a particular phosphorylation site. A binomial probability distribution is then used to calculate the location probability.

Since ion pairs matching candidate phosphorylation sites are considered, rather than the number of individual fragment ions that match an entire theoretical spectrum, the localization probability score is independent of the peptide identification probability score. If the peptide has a phosphorylated amino acid, the preliminary localization probability score is:

$$P = \sum_{i=0}^k \binom{4}{i} p^i (1-p)^{4-i} \quad (2)$$

where p is the probability for generating an ion pair randomly. Only MS2/MS3 b and y ion pairs are considered.

Within a given peptide, there may be several candidate phosphorylation sites. It is assumed that these are independent events. The final localization probability is as follows:

$$P=1 - \prod_i (1 - p_i) \quad (3)$$

where p_i is the probability of each phosphorylated amino acid assignment.

After using PepHMM to generate independent peptide identification scores for the MS2 and MS3 spectra and generating a localization probability score for the MS2/MS3 spectra pair, as described above, the three probabilities are combined into a final score for each possible peptide sequence and phosphorylation site localization. The top final score from among the various possibilities is viewed as an indicator of correct identification of both amino acid sequence and phosphorylation site(s).

PhosphoScan Correctly Localizes Phosphorylation Sites of an Affinity-Purified Hyperphosphorylated Protein

Using the PepHMM program, identification scores were ascertained for the MS2 and MS3 spectra in each data set. Furthermore, a localization probability for each possible set of phosphorylation sites was obtained according to the method outlined above. A final score that is comprised of a combination of all three scores was then obtained.

It is reasonable to expect that multiple instances of spectra for the same peptides will appear in a given set of spectra from the analysis of a sample. Thus, one would find a number of identical sequence/phosphorylation site matches for the correct identification of a phosphopeptide. We tested this assumption using a data set generated by Cripps et al.,¹⁶ which resulted from the analysis of the hyperphosphorylated protein, paired helical filament PHF-Tau. In this study, there were several doubly or singly phosphorylated peptides for which the phosphorylation sites could not be unambiguously localized by the SEQUEST program (Table 1), since the scores generated by SEQUEST alone were insufficient to differentiate the correct phosphorylation sites from among multiple serine and/or threonine residues. Using our PhosphoScan program, we successfully matched 61 of 63 phosphopeptide spectral pairs and confidently identified their phosphorylation site localizations; 59 of 63 peptides were doubly phosphorylated. Of the two remaining unmatched spectral pairs, for one of the pairs the MS2, MS3, and location scores for the various phosphorylation site options were all similar, whereas for the other, the three types of scores conflicted to the extent that no option was clearly superior. According to our observations, spectra for which PhosphoScan correctly localized phosphorylation sites may be classified into three cases, as shown in Table 2. In the first case, the MS3 spectrum is of a better quality than the MS2 spectrum. In the second case, the MS2 spectrum is of a better quality than the MS3 spectrum. In the third case, although the MS2 and MS3 spectra share similar scores, the phosphorylation site location probability is able to discriminate between them.

Location probability is calculated based on information obtained from the MS2/3 spectra by utilizing the product ions that precede and follow the phosphorylated amino acid. As it is rare for MS2 and MS3 spectra to contain all the pairs that may theoretically result from peptide fragmentations, the location probability is usually much less than 1 in most

predictions. However, since the probability rankings may be utilized to distinguish from among several predictions, use of the combined MS2/3 spectra may improve sequence and phosphorylation site identification.

Phosphopeptide Identification and Phosphorylation Site Localizations Are Greatly Improved by Using PhosphoScan as Opposed to SEQUEST and MASCOT

The examples shown in Table 2 demonstrate that PhosphoScan was able to correctly localize phosphorylation sites under the assumption that a correctly identified and localized peptide should appear multiple times in a sample. Synthesis and analysis of a phosphopeptide for which the sequence and phosphorylation sites are known should make it possible to check the correctness of predictions without depending on this assumption. To validate our program results, we tested PhosphoScan using the synthetic peptide STGSpATSLASpQGER.

Three data sets were generated under different conditions. In the first, an HPLC gradient method was used to introduce the peptide into a linear ion trap mass spectrometer; the other two were generated by direct infusion of the peptide into the mass spectrometer at different concentrations (see Materials and Methods). All spectra were generated using an instrument method that triggered a data-dependent MS3 scan upon detection of an appropriate neutral loss. Because all resulting spectra were generated from a single known peptide, we hypothesized that it would be possible to objectively determine which predictions of peptide sequences and phosphorylation sites were correct.

To ensure an unbiased comparison, we searched an in-house human database (compiled based on the Mascot MSDB database) using MASCOT, SEQUEST, and PhosphoScan with the same parameters: a peptide mass tolerance of 1.0 Da, trypsin digestion, PTM parameters of +80 and -18 Da on serine and threonine, and 1 missed cleavage allowed. Figure 5 compares the performance of the different programs in processing the MS2/3 phosphorylation spectra data sets. In this comparison, PhosphoScan overall achieved 80% correct phosphopeptide identifications while SEQUEST achieved 49% and MASCOT 51%. One drawback of combining MS2 and MS3 spectral data in phosphorylation site localization studies is that this approach does not increase the number of identified phosphopeptides, although it can improve phosphopeptide probabilities. In order to check if combining MS2 and MS3 spectra can generate more correct phosphopeptide predictions, we performed the following studies. Among these results, with sample 2 it was found that PhosphoScan correctly identified 15 from among 56 pairs for which SEQUEST assigned incorrect amino acid sequences for both the MS2 and the MS3 spectra, while with sample 3, PhosphoScan correctly identified 15 from among 26 such pairs mistakenly identified by SEQUEST. From among the relatively few spectral pairs available in sample 1, no pairs were found for which both the MS2 and MS3 spectra were incorrectly identified by SEQUEST. From these results, we conclude that PhosphoScan can correctly identify more phosphopeptides than either MASCOT or SEQUEST. MS2 and MS3 dependency information played an important role in increasing the number of correct phosphopeptide predictions. Correlation of the MS2 spectrum with the MS3 spectrum compensated for the cases where the spectra were not of

good quality. The localization probability was also of use in contributing to correct phosphorylation site localization.

In the data sets used in this study, it was noted that the phosphopeptide spectra were characterized by fragmentation patterns that were distinct from those of unmodified peptides. In the MS2 spectra of phosphopeptides, in addition to *b* and *y* ions, two other types of fragment ions were also frequently present: *b* and *y* ions with loss of water, and *b* and *y* ions with a neutral loss corresponding to loss of phosphoric acid. As indicated in Figure 6, in addition to a neutral loss of phosphoric acid precursor ion, several *b* and *y* neutral loss ions were also matched to the candidate peptide sequence. A small number of *b* and *y* loss of water ions were also identified (data not shown).

Discussion

By employing data-dependent neutral-loss MS3 scans and utilizing the dependency of MS2/MS3 spectra pairs on the detection of the loss of phosphoric acid, we were able to not only obtain peptide sequence identification, but also to identify the correct location of phosphorylation sites on hard-to-identify doubly phosphorylated peptides. While misidentification of phosphopeptide sequences and phosphorylation locations is not an infrequent occurrence with approaches limited to the evaluation of a single spectrum (MS2 or MS3), correlation of the MS2/3 spectral pair information decreases the likelihood that the same incorrect peptide match will be obtained for both spectra. Other advantages of our approach are increased database search speed and the ability to determine phosphorylation sites automatically without the need for manual validation.

A program that takes advantage of the unique fragmentation patterns of phosphopeptides while performing database searches might further enhance the accuracy of phosphopeptide scores. It was noted that with some doubly phosphorylated peptides, neutral loss ions were seen in the MS3 spectra corresponding to a second loss of phosphoric acid from the original phosphopeptide precursor ion. It is possible that a further data-dependent neutral loss scan could be performed on these ions; the resulting MS4 spectra could then be utilized for further confirmation of sequence and phosphorylation site location.

Several approaches to peptide phosphorylation site identifications have been reported. Ascore, developed by Beausoleil et al.,⁹ is applicable for general phosphorylation site localization based on MS2 data (<http://ascor.med.harvard.edu/ascor.php>). Lu et al.¹⁰ have developed an SVM approach to differentiate spectra of phosphopeptides from those of nonphosphorylated peptides. Our program, PhosphoScan, utilizes MS2/3 spectra pairs to identify true phosphorylation sites. Therefore, each program has its own advantages in addressing a specific aspect(s) of phosphorylation analysis. These programs are still in the stage of development; therefore, a direct comparison of these phosphorylation localization strategies is beyond the scope of the current study. At the same time, phosphopeptide data sources are quite limited and, moreover, are yet to be validated. Correct identification of phosphopeptides remains a challenging issue for most proteomics laboratories.

At this time, PhosphoScan supports up to two phosphorylation sites per peptide sequence. This method is based upon the presence of a neutral-loss precursor ion in the MS2 spectrum and on MS2/3 spectral pair information. If a real neutral loss peak is not observed in the MS2 spectrum of a phosphopeptide, and a corresponding MS3 scan is not triggered accordingly, the corresponding phosphopeptide will be missed. The source code of PhosphoScan is available upon request. In the future, PhosphoScan will be expanded so as to directly incorporate SEQUEST or MASCOT scores.

Acknowledgments

We thank Dr. Peter Gutierrez for method discussion and comments. This work is funded by National Institutes of Health Grants MH59786 and AG25323 to A.Y.

References

1. MacCoss MJ, Wu CC, Yates JR. *Anal Chem.* 2002; 74:5593–5599. [PubMed: 12433093]
2. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. *Electrophoresis.* 1999; 20:3551–3567. [PubMed: 10612281]
3. Mann M, Jensen ON. *Nat Biotechnol.* 2003; 21:255–261. [PubMed: 12610572]
4. Aebersold R, Mann M. *Nature.* 2003; 422:198–207. [PubMed: 12634793]
5. Steen H, Jebanathirajah JA, Rush J, Morrice N, Kirschner MW. *Mol Cell Proteomics.* 2006; 5:172–181. [PubMed: 16204703]
6. Diella F, Cameron S, Gemund C, Linding R, Via A, Kuster B, Sicheritz-Ponten T, Blom N, Gibson TJ. *BMC Bioinformatics.* 2004; 5:79. [PubMed: 15212693]
7. Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B. *Proteomics.* 2004; 4:1551–1561. [PubMed: 15174125]
8. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M. *Cell.* 2006; 127:635–648. [PubMed: 17081983]
9. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. *Nat Biotechnol.* 2006; 24:1285–1292. [PubMed: 16964243]
10. Lu B, Ruse C, Xu T, Park SK, Yates J. *Anal Chem.* 2007; 79:1301–1310. [PubMed: 17297928]
11. Zhang Z, McElvain JS. *Anal Chem.* 2000; 72:2337–2350. [PubMed: 10857603]
12. Ulintz PJ, Bodenmiller B, Andrews PC, Aebersold R, Nesvizhskii AI. *Mol Cell Proteomics.* 2008; 7:1–87. [PubMed: 17872894]
13. Jicha GA, Bowser R, Kazam IG, Davies P. *J Neurosci Res.* 1997; 48:128–132. [PubMed: 9130141]
14. Beer I, Barnea E, Ziv T, Admon A. *Proteomics.* 2004; 4:950–960. [PubMed: 15048977]
15. Wan Y, Yang A, Chen T. *Anal Chem.* 2006; 78:432–437. [PubMed: 16408924]
16. Cripps D, Thomas SN, Jeng Y, Yang F, Davies P, Yang AJ. *J Biol Chem.* 2006; 281:10825–10838. [PubMed: 16443603]

Figure 1.
Workflow for phosphopeptide identification.

Figure 2.

Spectra from true MS2/3 pairs are more similar than random pairs. A cosine distance score was used to evaluate the similarity between MS2/3 spectra pairs in our PHF-Tau data set. The histogram on the left represents the distribution of distance scores between random MS2/3 spectra pairs whereas the histogram on the right represents the distribution of distance scores between true MS2/3 spectral pairs.

Figure 3. Framework for predicting peptide sequences with phosphorylation options.

	B ion(MS2)	B ion(MS3)	Y ion(MS2)	Y ion(MS3)
T			1125.47	1027.47
P	199.11	<u>199.11</u>]	1028.42	<u>930.42</u>]
S*@	366.11	<u>268.11</u>]	<u>861.42</u>]	<u>861.42</u>]
L	479.19	<u>381.19</u>]	<u>748.34</u>]	<u>748.34</u>]
P	576.24	<u>478.24</u>]	<u>651.29</u>]	<u>651.29</u>]
T*	<u>757.26</u>	<u>659.26</u>]	<u>470.27</u>]	<u>470.27</u>]
P	854.31	756.31	373.27	373.27
P	951.36	853.36	276.17	276.17
T	1052.41	954.41	175.12	175.12
R				

$$\Pr(\text{first location is S}) = \sum_{i=0}^2 \binom{4}{i} p^i (1-p)^{4-i} = p_1$$

$$\Pr(\text{second location is T}) = \sum_{i=0}^3 \binom{4}{i} p^i (1-p)^{4-i} = p_2$$

$$P_{\text{location}} = 1 - (1 - p_1)(1 - p_2)$$

Figure 4.

Example of calculating location probability. An MS2/3 spectral pair was predicted as the phosphopeptide TPS_pLPT_pPPTR. (* indicates +80 modification, whereas @ indicates -18 modification.) A theoretical MS2/3 fragment ion table is shown above with matched fragment ions in bold. Fragment ion pairs located before and after phospho-sites are denoted by links between groups of such ions indicated with vertical lines. From this table, 2 out of 4 MS2/3 pairs were identified for the first phospho-site whereas 3 out of 4 MS2/3 pairs were identified for the second. Location probabilities were calculated based on these fragment ion pairs and the final probability was calculated by combining the two phosphosite probabilities.

Figure 5.

Phosphopeptide prediction results by SEQUEST, MASCOT, and Phosphoscan. Use of PhosphoScan demonstrated an improvement in identifying phosphorylation sites. The left-most set of bars shows percentages of incorrect predictions of both sequence and phospho-sites. The middle set shows percentages of correct sequence identification with incorrect phospho-site location assignments. The right set shows percentages of correct sequence identification with correct phospho-site location assignments. STGSpATSLASpQGER synthetic peptide was analyzed by LC-MS/MS/MS (Sample 1; 2 pmol) and by direct infusion into the mass spectrometer (Samples 2 and 3, 33 and 3.3 pmol/ μ L, respectively).

Figure 6.

Sequence and phospho-site identification for a custom peptide using MS2/3 coupling. SEQUEST was unable to correctly localize the phospho-sites of the synthetic phosphopeptide STGSpATSLASpQGER (SEQUEST could not distinguish STpG-SATSLASpQGER from STGSpATSLASpQGER). MS2 (A) and MS3 (C) spectra of the synthetic phosphopeptide are shown above with MS2 and MS3 b and y ion fragment masses shown in tables (B) and (D), respectively. The matched fragment ions are in bold and italics. In the MS2 spectrum, two serine residues have +80 modifications while in the MS3 spectrum, the first serine has a -18 modification, whereas the second has a +80 modification. A final probability was compiled from the combined probabilities. The table (E) generated by the PhosphoScan program clearly discriminates between possible phospho-sites using this final probability.

Table 1
Peptides from PHF-Tau Phosphopeptide Data Set with Phosphorylation Sites That Could Not Be Unambiguously Identified Due to Multiple Possible Phosphorylation Site Options Assigned by SEQUEST

residues ^a	peptide	phosphorylation options identified by SEQUEST
210–221	S _p RT _p PS _p LPT _p PPT _p R	Doubly phosphorylated on two of Ser210, Thr212, Ser214, Thr217 and Thr220
212–221	T _p PS _p LPT _p PPTR	Thr212 and Thr217 or Ser214 and Thr217
226–240	VAVVRT _p PPKS _p PS _p S _p AK	Thr231 and one of Ser235, Ser237, or Ser238

^aResidues indicate location on tau isoform that corresponds to the NCBI accession number NP_005901.

Table 2
Three Cases in Which Phospho-Site Identifications May Be Improved by PhosphoScan^a

MS2 sequence	PepHMM score	MS3 sequence	PepHMM score	location probability
Case 1: The MS3 spectrum is of a better quality than the MS2 spectrum				
SRT*PSLPT*PPTR	0.17	SRT#PSLPT*PPTR	0.91	0.20
SRTPS*LPT*PPTR	0.11	SRTPS#LPT*PPTR	0.72	0.05
Case 2: The MS2 spectrum is of a better quality than the MS3 spectrum				
T*PSLTPPT*R	0.56	T#PSLTPPT*R	0.99	0.05
T*PSLPT*PPTR	0.08	T#PSLPT*PPTR	1	0.03
Case 3: The MS2 and MS3 spectra share similar scores while the location probability discriminates between them				
SRT*PSLPT*PPTR	0.99	SRT#PSLPT*PPTR	0.99	0.20
SRT*PSLTPPT*R	0.98	SRT#PSLTPPT*R	0.99	0.05
S*RTPSLTPPT*R	0.84	S#RTPSLTPPT*R	0.99	0.05

^aThe first and second columns show peptide sequences and the corresponding scores predicted based upon the MS2 spectrum; the third and fourth columns show peptide sequences and corresponding scores predicted based upon the MS3 spectrum; the final column indicates the probability assigned to the phosphorylation locations.

indicates -18 modification of S or T.

* indicates +80 modifications of S or T.