

## PhosTShunter: A Fast and Reliable Tool to Detect Phosphorylated Peptides in Liquid Chromatography Fourier Transform Tandem Mass Spectrometry Data Sets

Thomas Köcher,\* Mikhail M. Savitski, Michael L. Nielsen, and Roman A. Zubarev

*Laboratory for Biological and Medical Mass Spectrometry, Uppsala University, Uppsala, SE-75123, Sweden*

Received November 7, 2005

A database independent search algorithm for the detection of phosphopeptides is described. The program interrogates the tandem mass spectra of LC-MS/MS data sets regarding the presence of phosphorylation specific signatures. To achieve maximum informational content, the complementary fragmentation techniques electron capture dissociation (ECD) and collisionally activated dissociation (CAD) are used independently for peptide fragmentation. Several criteria characteristic for peptides phosphorylated on either serine or threonine residues were evaluated. The final algorithm searches for product ions generated by either the neutral loss of phosphoric acid or the combined neutral loss of phosphoric acid and water. Various peptide mixtures were used to evaluate the program. False positive results were not observed because the program utilizes the parts-per-million mass accuracy of Fourier transform ion cyclotron resonance mass spectrometry. Additionally, false negative results were not generated owing to the high sensitivity of the chosen criteria. The limitations of database dependent data interpretation tools are discussed and the potential of the novel algorithm to overcome these limitations is illustrated.

**Keywords:** phosphorylation • mass spectrometry • phosphopeptides • identification • FT-ICR

### Introduction

The reversible phosphorylation of serine, threonine and tyrosine residues within proteins is one of the most important post-translational modifications (PTMs). It is estimated that in higher organisms more than 25% of all proteins are phosphorylated and more than 2% of the human genes code for protein kinases and the respective phosphatases.<sup>1</sup> Widespread diseases such as cancer<sup>2</sup> or neurological<sup>3</sup> and autoimmune disorders<sup>4</sup> can arise from aberrant protein phosphorylation. Protein stability and structure, protein localization and protein-protein interactions are partially regulated by this reversible modification, with the responsible kinases and phosphatases often organized in complex regulatory networks.<sup>5</sup> Protein phosphorylation is implicated to play a major role in many important cellular processes such as the cell cycle,<sup>6</sup> cell differentiation,<sup>7</sup> metabolism,<sup>8</sup> cell motility,<sup>9</sup> and signaling.<sup>10</sup> The identification of the phosphorylation sites of the involved proteins is critical for the understanding of these processes. Many of the substrates, kinases and phosphatases for the phosphorylation of histidine, lysine, arginine, aspartic acid, and glutamic acid residues have been found, but most of the biological and analytical literature deals with *O*-phosphorylations. The other types of phosphorylation are believed to be of lesser importance in higher organisms and are also unstable under the conditions of standard sample preparation tech-

niques. This work exclusively targets the phosphorylation of serine and threonine, together accounting for at least 90% of all phosphorylated amino acids in higher organisms.

Mass spectrometry has become the method of choice for the identification of proteins and their PTMs.<sup>11-14</sup> However, while the identification of proteins can be considered as a routine task with the current technology, the complete mapping of the PTMs even in a single protein is still a challenging task. There are multiple problems involved in the mass spectrometry-based analysis of phosphorylation sites, occurring at all steps of the analysis. Some of them are typical for all PTMs but labile PTMs such as phosphorylations or glycosylations are especially difficult to analyze. The analysis of phosphorylation sites usually follows the general principles developed for protein identification.<sup>15</sup> These protocols start with the digestion of the proteins of interest with one or more proteases such as trypsin, followed by the analysis of the resultant peptide mixture by mass spectrometry. In most cases, tandem mass spectrometry is used for peptide sequencing. The spectra are interpreted either manually or the combined data set is searched against a protein database by one of the various search programs.<sup>16,17</sup> The fundamental complication in the analysis of modified proteins is that whereas any two tandem mass spectra of typical tryptic peptides unambiguously identify a protein in a protein database, for a complete PTM analysis all modified peptides have to be detected and sequenced. Consequently, either the specific detection of all modified peptides or complete sequence coverage is required, both hard to achieve in practice.

\* To whom correspondence should be addressed. Tel: +46-18-4715729. Fax: +46-18-4715729. E-mail: Thomas.Kocher@bmms.uu.se.

When only one proteolytic enzyme is used, some of the generated peptides might have high molecular masses, and thus are less suited for standard mass spectrometry routines and on-line high performance liquid chromatography (HPLC). Additionally, the physicochemical properties of very acidic or hydrophilic peptides are disadvantageous for the complete MS analysis including sample preparation. Hydrophilic peptides might be lost during the initial desalting steps of the analysis. Small or very acidic peptides are observed preferentially as singly charged ions, which are in general less suited for tandem mass spectrometry but are particularly futile for electron capture dissociation (ECD).<sup>18</sup> The properties of the PTM itself might be problematic, such as the high binding affinity of the phosphate group of a phosphorylated peptide to the metal parts of the HPLC system. The heterogeneity of modified proteins is another frequently encountered issue, leading to a lower abundance and a reduced detection probability of each individual modified peptide.

Apart from their detection, peptides phosphorylated on serine or threonine residues impose additional problems for tandem mass spectrometry because of the labile nature of their phosphate moiety. When activated by collisionally activated dissociation (CAD) or related techniques, phosphoric acid is lost as a neutral species from both the precursor ion and the fragment ions. Novel tandem mass spectrometry techniques such as ECD,<sup>18,19</sup> or the recently discovered method of electron-transfer dissociation<sup>20</sup> (ETD) offer the advantage that otherwise labile modifications such as phosphorylations<sup>21</sup> or glycosylations<sup>22</sup> are stable during fragmentation. On the other hand, the high fragmentation efficiency of labile modifications offers a means for their specific and sensitive detection. Neutral loss scanning<sup>23</sup> and precursor ion scanning<sup>24</sup> have both been successfully applied to the analysis of phosphorylation sites. Precursor ion scanning in the negative ion mode for the phosphorylation specific product ion of  $m/z$  -79 is a very sensitive and selective technique usually performed on triple quadrupole instruments.<sup>24,25</sup> The method is handicapped by the necessary toggling between the negative and the positive ion mode and its dependence on buffers with a high pH, making the technique less useful for hyphenated methods. Precursor ion scans for cationic marker ions such as the immonium ion of phosphorylated tyrosine have been applied to other mass spectrometers such as quadrupole time-of-flight instruments (qTOFs).<sup>26,27</sup> In neutral loss scanning, the loss of phosphoric acid is monitored,<sup>28</sup> and, as a result, phosphorylated tyrosine residues are not targeted by this method. Similar to precursor ion scanning, neutral loss scanning can be implemented in the most simple and straightforward way on triple quadrupole instruments but the method is used with other instrumentation such as qTOFs.<sup>29,30</sup> The drawbacks associated with this method are its charge state dependence and its low selectivity due to competing fragmentation channels.<sup>29,31</sup> There are several biochemical approaches to the analysis of phosphorylated proteins such as the specific enrichment of phosphopeptides by metal-based affinity purification (IMAC)<sup>31,32</sup> or the specific labeling of phosphorylated peptides.<sup>33</sup> Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) instrumentation is usually used for the differential analysis of protein digests before and after phosphatase treatment.<sup>34</sup> Naturally, the most straightforward approach to the analysis of phosphorylated proteins is the analysis of all proteolytic peptides by on-line HPLC-MS/MS in a data dependent acquisition mode.<sup>35</sup> This is the method of choice if additional modifications of a protein should

be detected. However, a two-dimensional separation strategy might be necessary such as a combination of reversed phase chromatography with IMAC<sup>36</sup> or ion-exchange chromatography.<sup>37,38</sup>

Regardless of the applied method, the detected phosphopeptides have to be assigned to the sequence of the protein. Even if all modified peptides are detected and subjected to tandem mass spectrometry, the identification and the localization of labile PTMs to specific amino acids can be tedious or even impossible. First, the phosphorylation site is not necessarily covered by the observed fragment ions, especially in tandem mass spectra of peptides with high molecular weights. Second, the obtained spectra will be more difficult to interpret because of the additional peaks corresponding to the neutral losses in the spectrum. The fragment ions still bearing the intact modification might even be absent. Moreover, additional and consecutive losses will be observed in the product ion spectrum if the phosphorylated peptide contains either additional labile modifications such as oxidized methionine, glycosylated amino acids, or multiple phosphorylation sites. Concurrent fragmentation of multiple precursors in a single MS/MS experiment complicates the interpretation even further.<sup>32</sup> In a manual analysis technique, such as nanoelectrospray-based precursor ion scanning, the interpretation step is performed by the analyst.<sup>24</sup> The identification of the modified peptide can be aided by error-tolerant search programs.<sup>39</sup> The manual data interpretation is practical because only a relatively small number of spectra have to be interpreted. In contrast, typical high throughput approaches such as LC-MS/MS experiments<sup>36</sup> generate hundreds of tandem mass spectra in a single run. The manual interpretation of the data sets is not feasible and the combined data set has to be interrogated by data interpretation tools such as Mascot<sup>16</sup> or SEQUEST.<sup>17</sup> However, some product ion spectra of phosphopeptides might not be of sufficient quality for a successful identification with a search engine. Additionally, the phosphopeptides might bear an additional unexpected modification, thus escaping identification. Although all search engines allow the consideration of fixed or variable modifications it is not possible to take all potential modifications into account. Even a single variable modification such as phosphorylation greatly increases the complexity of a search, resulting in much longer search times and reduced specificity.

Here we describe a fast and reliable data mining algorithm, capable of locating the spectra of phosphorylated peptides in LC-MS/MS data files circumventing the loss of valuable data. Our program, called PhosTShunter, aims to detect phosphopeptides in a database independent manner in Fourier transform ion cyclotron resonance (FT-ICR) MS/MS data sets associated with the highest quality data available today in mass spectrometry. FT-ICR MS ensures high resolution and parts-per-million mass accuracy of molecular mass determination.<sup>40</sup> Two complementary fragmentation techniques, CAD and ECD, are employed. Their combined use is shown to provide a wealth of structural information on every detected peptide, which greatly improves the efficiency and validity of database searches and allows for high-throughput de novo sequencing of unmodified peptides.<sup>41</sup> The experiment is not designed for any modification in particular, ensuring that the algorithm can be applied a posteriori to any previously generated set of data. Second, the new approach is database independent, and can be applied to peptides from organisms with unsequenced genomes. PhosTShunter does not require extensive MS/MS

information using even MS/MS data without peptide backbone fragmentation. Defining specificity by the ratio between the number of true negative hits and the number of all spectra from unphosphorylated peptides, the criteria based on the neutral loss of phosphoric acid from the precursor ion proved to be highly specific when applied to FT-ICR MS/MS data sets by practically excluding false positive hits. Below we describe and validate PhosTShunter and compare it with the output of the Mascot search engine which was used for identification of the analyzed proteins and a subset of phosphopeptides. PhosTShunter was expected to find and confirm all phosphopeptides that Mascot was able to identify and on top of that detecting new phosphopeptides, if present. Every additionally detected peptide was inspected and validated manually. The rate of false positives and false negatives in the PhosTShunter output was tested on a set of known samples.

## Materials and Methods

**Sample Preparation.** Proteins were reduced 20 min with 10 mM DTT (Sigma; St. Louis, MO) in 50 mM ammonium bicarbonate at 56 °C followed by reaction with iodoacetamide (Sigma; St. Louis, MO) at a final concentration of 100 mM in the dark for 25 min at room temperature. The tryptic digestion was performed in 50 mM ammonium bicarbonate with a protein concentration of 10  $\mu$ M at 37 °C using trypsin (Promega; Madison, WI) at a final concentration of 12 ng/ $\mu$ L. Bovine serum albumin and chicken ovalbumin were bought from Sigma (Sigma; St. Louis, MO). Synthetic peptides were synthesized in-house using a ResPepMicroScale peptide synthesizer (INTAVIS Bioanalytical Instruments AG; Köln, Germany) and purified on a preparative HPLC system.

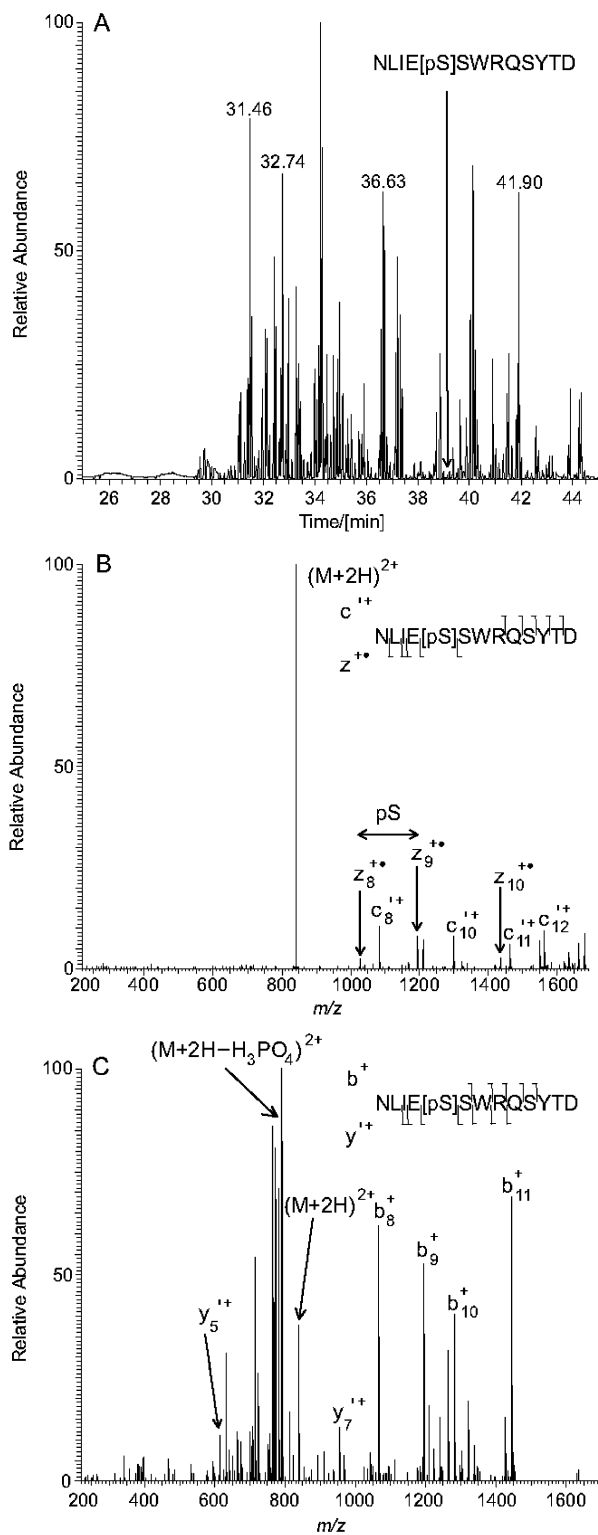
**Mass Spectrometry.** All experiments were performed on a 7-tesla hybrid linear ion trap (LTQ FT) mass spectrometer (Thermo Electron; Bremen, Germany) operated with a nano-electrospray ion source (Proxeon Biosystems; Odense, Denmark). All solvents used were of the highest purity available (Merck; Darmstadt, Germany). The high performance liquid chromatography setup used in conjunction with the mass spectrometer was an Agilent 1100 nanoflow system (Agilent; Palo Alto, CA) and consisted of a solvent degasser, a nanoflow pump, and a thermostated microautosampler. A 15-cm fused silica emitter with a 75- $\mu$ m inner diameter and a 375- $\mu$ m outer diameter (Proxeon Biosystems; Odense, Denmark) was used as an analytical column. The emitter was packed in-house with a slurry of reverse-phased, fully end-capped Reprosil-Pur C18-AQ 3- $\mu$ m resin (Dr. Maisch GmbH; Ammerbuch-Entringen, Germany) dispersed in methanol using a pressurized "packing bomb" operated at 50–60 bar (Proxeon Biosystems; Odense, Denmark). Mobile phases consisted of A (0.5% acetic acid and 99.5% water (v/v)) and B (0.5% acetic acid and 10% water in 89.5% acetonitrile (v/v)). A five-microliter portion of prepared peptide mixture was automatically loaded onto the column and rinsed for 20 min in 2% buffer B at a flow rate of 500 nl/min followed by a 90-min gradient from 2 to 45% buffer B at a constant flow rate of 200 nl/min. The analysis was performed in a data-dependent acquisition mode. In this setting, the mass spectrometer automatically toggles between a survey MS scan and consecutive ECD and CAD MS/MS scans of the most abundant peptides detected eluting at that moment from the nano-LC column. The survey scan was performed in the FT cell recording in an  $m/z$  window between 300 and 1500 Th. The resolution was set to 100 000 and the automatic gain control (AGC) was set to 3 000 000 ions. If the ion of interest

was multiply charged, consecutive ECD (AGC: 900 000 ions) and CAD (AGC: 600 000 ions) MS/MS spectra were acquired with a resolution of 25 000 and an isolation window of  $m/z$  7. The irradiation time in ECD was 70 ms and the activation time in CAD 30 ms. The maximum accumulation time during CAD and ECD was 1s. The masses of fragmented ions were put on an exclusion list for 240 s.

**Data Analysis.** The original Xcalibur (Thermo Electron; Waltham, MA) binary data (RAW-files) were converted by BioWorks TurboSEQUENT (Thermo Electron; Waltham, MA) into a set of peak lists (DTA-files). These DTA-files contain the mass and the charge state of the precursor, as well as the  $m/z$  values and intensities of all the fragment ion peaks in the spectrum above a certain cutoff intensity value. All software for processing the DTA-files was developed in JAVA. Two DTA-files are present for each precursor ion, one generated by CAD and the other one by ECD fragmentation. In the first step, PhosTShunter processes the DTA files by deisotoping and charge-deconvoluting to the neutral state. Details of this procedure are published.<sup>41,42</sup> Then the program uses the information contained in the DTA-files of each precursor ion to determine whether it is reported as a phosphopeptide. This is done by testing if at least one criterion of an orthogonal set of criteria is fulfilled. The first criterion (C1) in this ensemble is defined by the presence of a fragment ion in the CAD DTA-file generated by the neutral loss of phosphoric acid. The second criterion (C2) is defined by the presence of a fragment ion generated by a combined neutral loss of phosphoric acid and water in the CAD DTA-file. Other criteria were tested (see results section) but finally dismissed. If one of the criteria is matched the program will report the molecular mass of the peptide, its file number and all fulfilled criteria. The interpretation of the found spectra was performed manually. The mass accuracy applied for the deisotoping of the spectra can be chosen by the user. A second value is set for the  $m/z$  window of the applied criteria. In our experimental setting these values were set to  $m/z$  0.01. The program will be accessible at <http://www.bmms.uu.se>.

## Results and Discussion

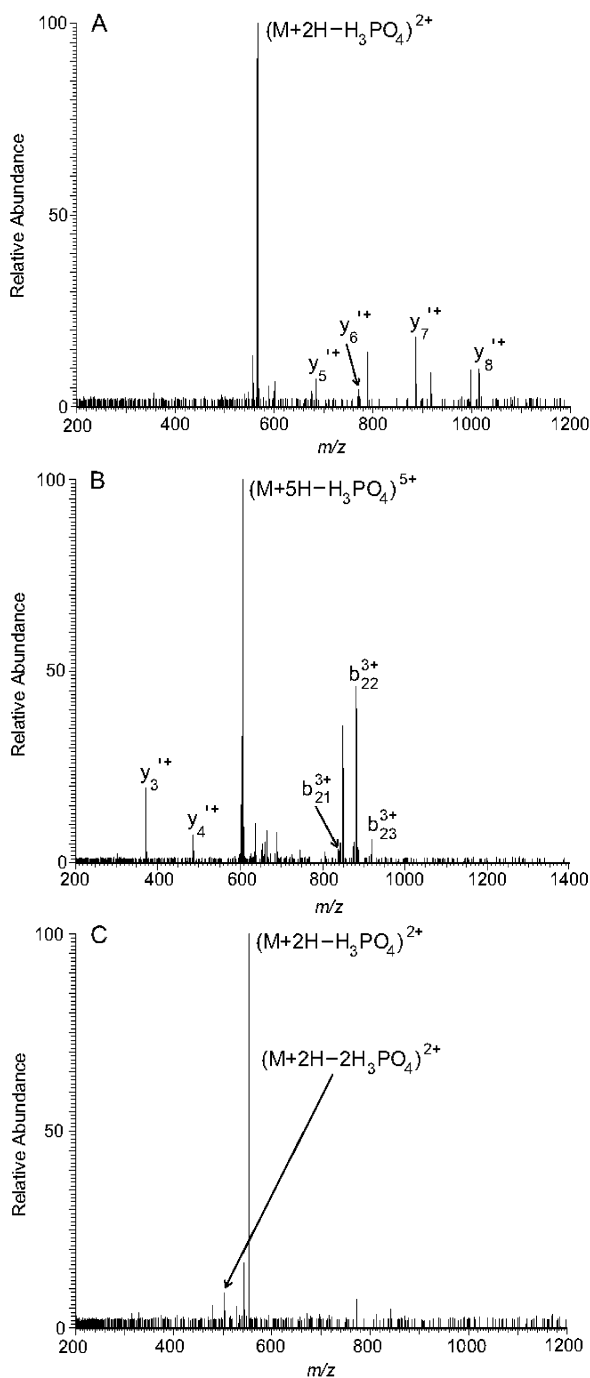
**Experimental Strategy.** We chose an experimental strategy based on the on-line HPLC separation of proteolytic peptides and data dependent tandem mass spectrometry using FT-ICR MS instrumentation. The achieved mass resolution of 100 000 allowed the automatic and correct assignment of the charge states and the monoisotopic peaks of phosphorylated peptides, up to masses of 5000 Da. In practice, we obtained mass accuracies well below two ppm, which is in our opinion crucial for a correct and unambiguous assignment of the modified peptides. Each detected peptide was picked for two subsequent MS/MS experiments. First, electron capture dissociation (Figure 1, Panel B) was applied, followed by an independent fragmentation using CAD (Figure 1, Panel C). In many cases, the sequence covered by the observed product ions increased by the use of both methods, such as in the spectra of the doubly charged ( $m/z$  839.8593) phosphopeptide NLIE[pS]SWRQSYTD (Figure 1, Panel B and Panel C). Additionally, ECD generates exclusively fragment ions still bearing the phosphate group while at the same time CAD provides us with data containing the indicative neutral loss of phosphoric acid (Figure 1, Panel C). The interrogation of MS/MS spectra by our algorithm is based on comparing experimentally obtained mass differences with theoretically calculated mass values, such as the mass



**Figure 1.** LC-MS/MS analysis of a tryptic digest of BSA spiked with the synthetic phosphopeptide NLIE[pS]SWRQSYTD. Panel A. In the total ion chromatogram the peak corresponding to the substoichiometric phosphopeptide is indicated. Panel B. The ECD spectrum of the peptide NLIE[pS]SWRQSYTD with the  $c^{1+}$ -ion and the  $z^{2+}$ -ion series labeled. Panel C. The CAD spectrum of the phosphopeptide with the observed cleavage sites assigned to the sequence. The  $y^{1+}$ -ion series, the  $b^{+}$ -ion series and the product ion corresponding to the neutral loss of phosphoric acid are indicated. The position of the phosphoserine residue was confirmed by the  $y_7^{1+}$ - and the  $y_8^{1+}$ -ion.

difference between two fragment ions in an ECD spectrum and the theoretical mass of phosphoserine. In this context, the high mass accuracy and the high mass resolution of the product ion spectra were again important in order to identify the monoisotopic peaks and to avoid false positive identifications.

**Limitations of Database Dependent Search Engines.** To investigate the required features for a novel algorithm and the shortcomings of current data interpretation tools for the analysis of phosphopeptides in the LC-MS/MS data files, the proteolytic digest of the highly modified protein osteopontin was analyzed. The protease GluC was used, cleaving after glutamic acid and to a lesser extent after aspartic acid. The search engine Mascot confidently identified osteopontin based on the LC-MS/MS data set with a Mascot score of 599 and found a total of 10 phosphorylation sites. Although Mascot identified 23 different phosphopeptides, it did not identify all phosphopeptides in the data set. When we interrogated the LC-MS/MS data with the algorithm described in the next section, 146 spectra of putative phosphopeptides and their salt adducts were found. Three main types of spectra detected by our algorithm but not identified by Mascot were observed. In the case of the peptide SQEDSKL[pS]QE (Figure 2, Panel A) the deamidation or mutation in the original sequence SQENSKLSQE from asparagine to aspartic acid prevented identification of the phosphopeptide by Mascot. A database dependent search engine inevitably fails to identify a peptide if its amino acid sequence is mutated or modified in an unexpected way even in the case of a highly informative spectrum. It should be mentioned that such peptides will be found by Mascot in an error-tolerant search<sup>43</sup> if the product ion spectra is of sufficient quality. Likewise, known additional modifications can be considered as variable modifications in the search parameters. In the case of the deamidated peptide SQEDSKL[pS]QE Mascot identified the phosphopeptide when deamidation was an allowed variable modification. However, it is not possible to take all potential modifications into account because of the exponentially growing time demands for the database search. In another type of spectrum (Figure 2, Panel B) the reason for the failed automatic annotation by Mascot was the poor informational content of the spectrum. However, the presence of three b-ions and two y-ions allowed the manual identification of the five times charged and doubly phosphorylated peptide MHDAPKKTSQLTDH[2pSEETNS]DELPKE ( $m/z$  626.2682) from osteopontin. The two phosphorylation sites could only be assigned to the C-terminal region of the peptide because of the weak sequence coverage in the corresponding part of the spectrum. It should be noted in the context of this example that Mascot does not consider fragment ions with a higher charge state than two and it does not report phosphopeptides when the phosphorylation site cannot be localized within the peptide. Searching the charge deconvoluted and deisotoped DTA files of osteopontin resulted in the detection of five additional spectra of phosphopeptides. However, the five times charged phosphopeptide MHDAPKKTSQLTDH[2pSEETNS]DELPKE was still not identified by Mascot. The spectrum of the doubly phosphorylated peptide SHHSDSEDE (Figure 2, Panel C) was an extreme case of a spectrum with poor informational content. We observed only product ions corresponding to the loss of phosphoric acid from the precursor ion in the CAD spectrum. Naturally, this peptide was not identified by Mascot. The identification of this phosphopeptide, however, was with high confidence based on the precise precursor ion mass recorded with the FT-ICR MS and the



**Figure 2.** Examples of phosphopeptides from osteopontin which were not identified by a database dependent search engine are shown. In all cases, the loss of phosphoric acid was the most dominant fragmentation channel. Panel A. The  $y^{+}$ -ion series of the peptide SQEDSKL[pS]QE is indicated. The CAD data allowed locating the phosphorylation site and the identification of the peptide. Panel B. The CAD spectrum of the five times charged and doubly phosphorylated peptide MHDAPKKTSQLTDH-[2pSEETNS]DELPKE is shown. In this case, the search engine failed because the phosphorylation sites cannot be precisely located and most of the fragment ions have a charge state higher than two. Panel C. The CAD spectrum of the doubly phosphorylated peptide SHHSDESDE typifies an extreme case of a product ion spectrum with poor data quality. The identification is based on the precise mass measurement of the peptide mass and product ions generated by consecutive losses of phosphoric acid.

evident single and double loss of phosphoric acid. The difference between the theoretical mass calculated from the sequence of osteopontin and the experimental mass was 4 mDa guaranteeing reliable sequence assignment in the case of a highly purified protein sample.

**Potential Criteria Characteristic for Phosphopeptides.** The desired features for the novel algorithm should be speed, reliability and database independence. It should be able to report all spectra of phosphopeptides in the data files of on-line HPLC FT-ICR MS/MS experiments without significant false positive or false negative results. In addition, it should be fast, finding the spectra of the phosphopeptides in a matter of minutes. We evaluated several potential criteria for the detection of phosphopeptides from LC-MS/MS data sets generated by ECD and CAD. Usually, the phosphopeptide specific neutral loss of phosphoric acid (97.977 Da) from the peptide ion is the most favored fragmentation reaction in CAD or related techniques<sup>44</sup> (Figure 2) and, therefore, was expected to be the most efficient way to recognize phosphopeptides.<sup>45</sup> The loss of phosphoric acid is sometimes accompanied by the loss of a water molecule leading to a neutral loss of 115.987 Da. We noticed in our experiments that in most product ion spectra both losses were present but some precursor ions in higher charge states preferentially lost water and phosphoric acid. Consequently, our first two criteria were defined by the requirement that there must be a peak in the charge-deconvoluted spectrum 97.977 Da (C1) or 115.987 Da (C2) lower in mass than the precursor ion. Fragmentation along the peptide bond can also coincide with a neutral loss of phosphoric acid. To exploit this reaction for the third criterion (C3), the program interrogated the CAD DTA-files for the presence of at least two pairs of fragment ions spaced by a mass of 97.977 Da. However, if the entire peptide fragment ion population has lost phosphoric acid, criterion C3 would not be applicable. In this case, the product ion series in the ECD spectrum would be unchanged in their  $m/z$  values whereas the complementary ion series in the CAD spectrum would be 97.977 Da lower in mass. To account for this possibility, we used the previously described golden complementary pair approach.<sup>41,46</sup> In the corresponding criterion (C4) at least two golden complementary pairs altered by a mass shift of 97.977 daltons must be present in a pair of ECD and CAD spectra from a single precursor. The algorithm identifies the corresponding ion pairs by their mass difference of  $-115.004$  Da in the case of the b- and the c-ion series or  $-81.959$  Da for the y- and the z-ion series. Again, two such pairs are required for a positive identification. The other set of criteria was based on the possibility that fragment ions are observed which have not lost phosphoric acid upon activation. In ECD, modifications are always retained but even CAD does not always induce the loss of phosphoric acid from the entire product ion population in the course of peptide chain fragmentation. The fifth criterion (C5) is based on the presence of two fragment ions matching to the mass of an intact phosphorylated amino acid (Figure 1, Panel B) in the ECD spectra. Similarly, the sixth criterion (C6) interrogates the CAD spectra for the presence of two fragments spaced by the mass of a phosphorylated serine or threonine.

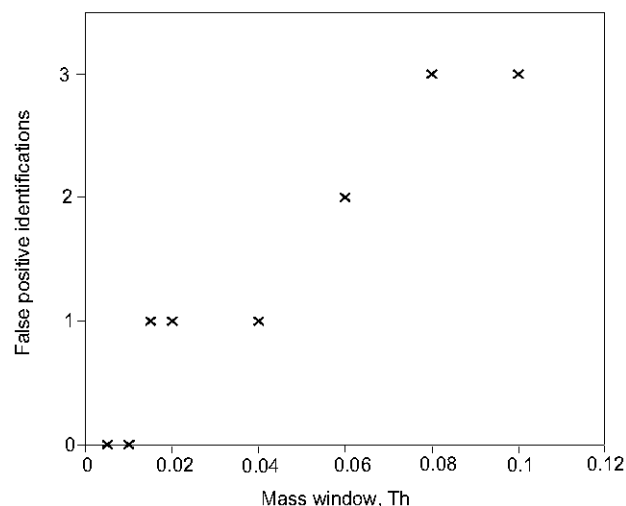
**Evaluation of the Suitability of Each Criterion.** The applicability of the different approaches was tested with 50 fmol of a peptide mixture obtained from an in-solution digest of bovine serum albumin (BSA) spiked with less than 5 fmol of the synthetic phosphopeptide NLIE[pS]SWRQSYTD. This model experiment mimicked the analysis of a phosphoprotein con-

taining a single and substoichiometric phosphorylation site. The peptide mixture was separated by nano-LC and analyzed on-line by FT-ICR MS/MS in a data dependent acquisition mode (Figure 1, Panel A). The two most intense precursor ions from each survey mass spectrum were selected for fragmentation. Each peptide was first fragmented by ECD (Figure 1, Panel B) followed by CAD (Figure 1, Panel C). The analysis of the combined DTA-files with the PhosTShunter algorithm was completed in less than a minute. Each of the six criteria identified the tandem mass spectra of a doubly charged peptide ( $m/z$  839.8593), the added phosphopeptide. However, applying criterion C5 led to 28 false positive results and criterion C6 incorrectly identified 35 mass spectra. The merged DTA-files were searched against the complete protein database with Mascot, allowing the variable modifications of oxidized methionine and the phosphorylation of serine or threonine. Mascot identified BSA and some contaminating proteins in the sample but the added phosphopeptide was not found because of its artificial sequence.

Without an algorithm such as PhosTShunter the identification of the synthetic phosphopeptide would require the manual inspection of the complete LC run. The latter can be aided by a neutral loss monitoring option of the Xcalibur software. However, this option was found to be associated with the shortcomings characteristic for neutral loss scanning. First, false positives were created when the charge state of the precursor ion and the potential neutral loss fragment ions did not match. Second, all charge states had to be searched separately. These inadequacies might be addressed by charge deconvolution and deisotoping of the raw data similar to the procedure used by PhosTShunter. Without processing of the raw data, these two issues inhibit the desired fast and straightforward reduction of the data set to a few spectra, preferentially to the single spectra corresponding to the phosphorylated peptide. In practice, the neutral loss monitoring option in the Xcalibur software detected four putative phosphopeptides when searching for a loss of  $m/z$  48.9 and only considering peaks above a threshold of five percent of the peak with the highest intensity. The spiked phosphorylated peptide was detected but the most prominent hit was a false positive result. The analysis of neutral losses from assumed higher charge states resulted in even higher numbers of false positives with 14 spectra for the loss of  $m/z$  32.66. In total 47 peaks were detected when analyzing the  $m/z$  losses corresponding to all charge states from doubly charged ions up to five times charged ions. In addition to the correct phosphopeptides, 46 false positive results were obtained from the 271 CAD spectra of the LC run. Consequently, even aided with a neutral loss screening feature manual inspection is slow and a tedious task. In contrast, PhosTShunter interrogating the complete set of spectra with the criteria C1 and C2 identified only the spiked phosphopeptide without any false positive results.

It should be stressed at this point that the algorithm itself obviously does not improve the initial detection of phosphopeptides by the mass spectrometer but it ensures that already detected and sequenced peptides are not lost during the data interpretation process. However, phosphorylated peptides, especially when present at substoichiometric amounts in complex mixtures, might not be selected for fragmentation and, therefore, escape identification.

Naturally, a more complex sample was needed to assess the false positive rate of the algorithm and each criterion used. We decided to analyze an in-solution digest of an *E. coli* whole

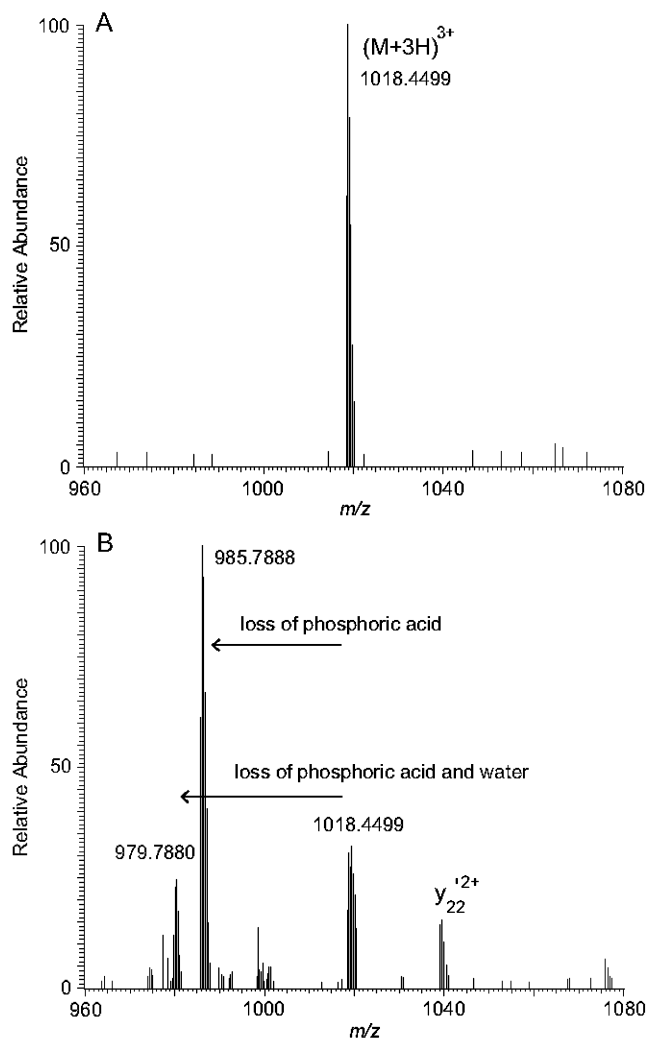


**Figure 3.** Assessment of the direct influence of mass accuracy on the detection of false positives in the proteolytic digest of an *E. coli* whole cell lysate. The mass window used for the deisotoping was left unchanged with an  $m/z$  value of 0.01. False positive results were observed starting with  $m/z$  0.015.

cell lysate as a negative control sample. In bacteria, protein phosphorylation of serine and threonine residues is present to a much lesser extent than in higher organisms. Therefore, this sample combined sufficient complexity with a low likelihood for the detection of phosphorylated peptides. The peptide mixture was analyzed with on-line HPLC FT-ICR MS/MS as described above. The data set contained 2587 pairs of product ion spectra. A Mascot search identified approximately 200 proteins. As expected, a database search with the possibility of phosphorylated serine or threonine did not identify any phosphorylated peptides. The data files were investigated with PhosTShunter and again applying criteria C1, C2 and C4 did not lead to the identification of any phosphorylated peptides. Applying criteria C3, C5 and C6 led to 11, 72 and 180 false positive results, respectively. The analysis of the data files with the algorithm was finished in less than a minute.

Next we further assessed the role of high mass accuracy for the criteria C1 and C2. The data files from the *E. coli* sample were analyzed with PhosTShunter, increasing stepwise the  $m/z$  windows of the applied criteria. With a window of  $m/z$  0.015 a single false positive result was obtained with criterion C1 (Figure 3) but none with a window of  $m/z$  0.01. Additionally, this experiment showed the robustness of the criteria C1 and C2 because even with large  $m/z$  windows the number of false positive results stayed surprisingly modest (Figure 3). We believe that this was due to the discriminating effect of the requirements that only equal charge states of the precursor and fragment ion were considered and that both peaks had to have an isotopic distribution. Summarizing the results from both experiments, the criteria C1, C2, and C4 identified the phosphopeptide in the background of the proteolytic peptide mixture and each of the three criteria was stringent enough to exclude false positive identifications in the complex proteolytic peptide mixture derived from the *E. coli* whole cell lysate. Additionally, the PhosTShunter algorithm was found to be fast, fulfilling the desired requirement of high-speed analysis.

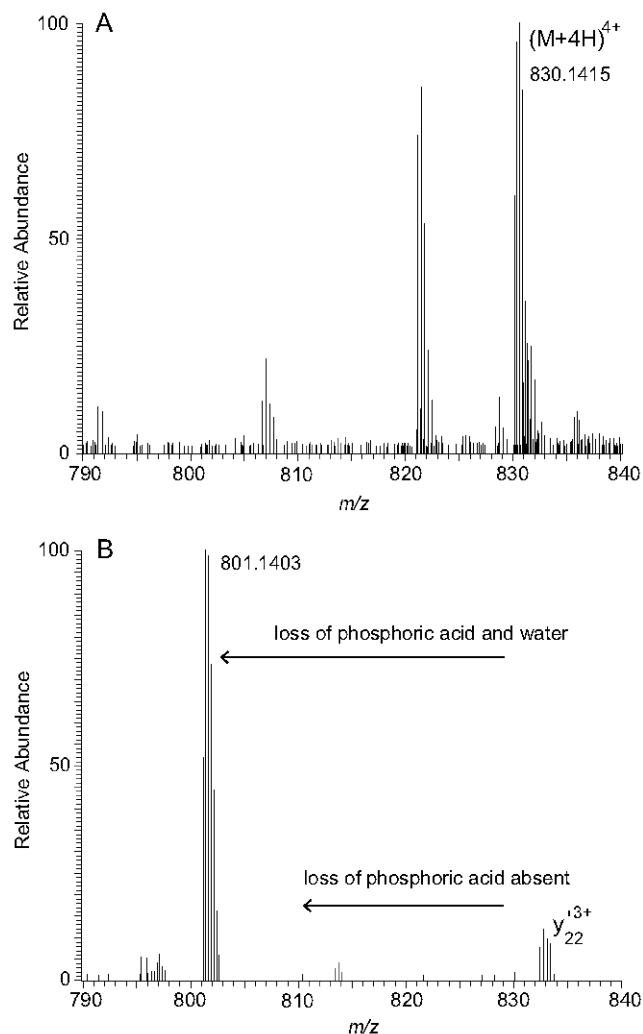
**Analysis of a Model Phosphoprotein.** To further evaluate the efficiency of PhosTShunter, the model phosphoprotein ovalbumin was analyzed. First, the protein was reduced, alkylated and digested with trypsin. The generated peptide



**Figure 4.** Expanded  $m/z$  view to illustrate the neutral losses observed. Panel A. The triply charged ion of the phosphopeptide  $m/z$  1018.4499 is labeled in the FT-ICR MS survey scan. Panel B. The loss of phosphoric acid and the loss of phosphoric acid and water are indicated in the CAD spectrum.

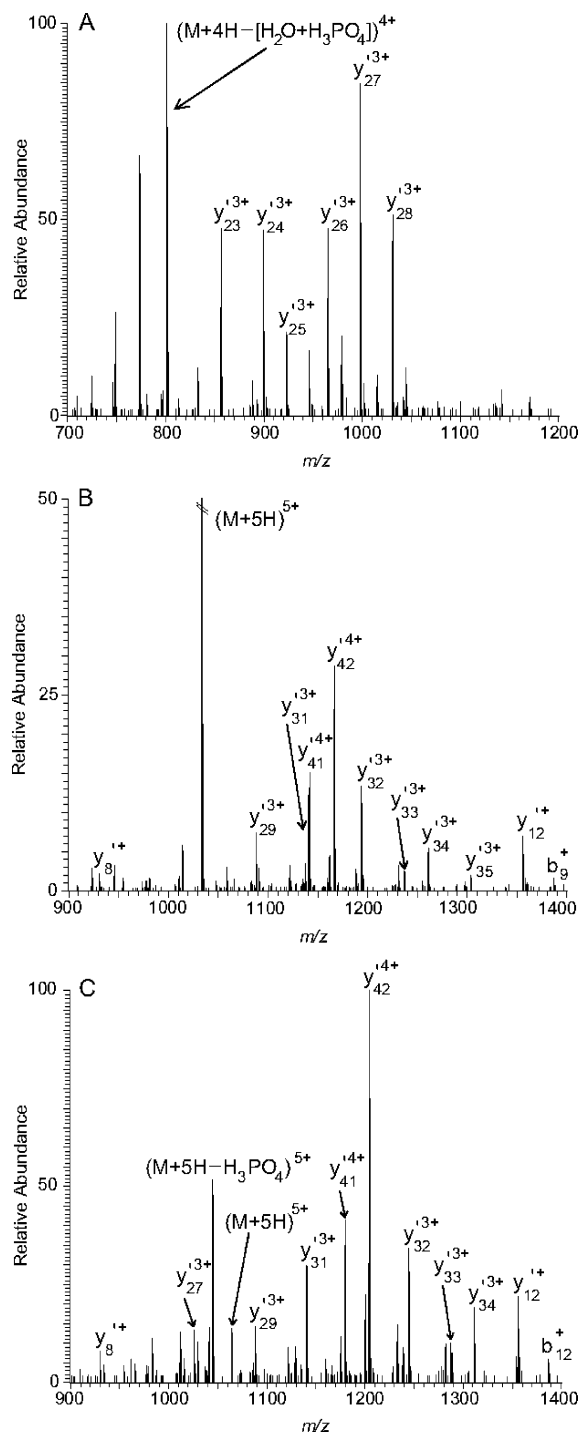
mixture was then analyzed by on-line HPLC FT-ICR MS/MS and the combined data files were submitted to Mascot with the search parameters set to a fixed modification of carboxyamidomethylation of cysteine and variable modifications of oxidized methionine and phosphorylation of serine and threonine. The protein ovalbumin was identified with a Mascot score of 449 and the two phosphopeptides EVVG[pS]AEAGVDAAS-VSEEF (  $m/z$  1044.9705) and FDKLPGFGD[pS]IEAQ[camC]-GTSVNVHSSLR (  $m/z$  726.0925) were identified.

In the next step, the DTA files were interrogated with our algorithm. Consistent with the previous experiments, additional 340 false positive results were obtained with the criteria C3, C5, and C6. Application of criterion C4 showed a poor performance by finding only a single phosphopeptide, which was also detected by criteria C1 and C2. A thorough inspection of the spectra revealed that the weak data quality of the ECD spectra was responsible for this failure. Consequently, at the end of our evaluation of the potential criteria only criteria C1 and C2 were left as the most sensitive and specific criteria, finding in total 18 phosphopeptides or their potassium adducts in the data files of the ovalbumin digest. The detection of the



**Figure 5.** Expanded section of the mass spectra of the phosphopeptide  $m/z$  830.1415. Panel A. The four times charged ion is labeled in the FT-ICR MS survey scan. Panel B. The combined loss of phosphoric acid and water is indicated in the CAD spectrum. The peaks corresponding to the precursor ion and to product ions generated by the loss of phosphoric acid from the precursor ion are absent.

potassium adducts of the phosphopeptides testified to the high sensitivity of the experimental part of our approach. In most cases, application of either C1 or C2 proved to be redundant (Figure 4) but five phosphopeptides were identified only by criterion C1 (Figure 5), while four phosphopeptides were identified only by criterion C2. In the spectra of two phosphopeptides, such as in the spectra of the five times charged peptide at  $m/z$  830.1422 (Figure 5), only the combined loss of phosphoric acid and water was observed, with the fragment ions generated by the neutral loss of phosphoric acid completely absent. In other spectra, the weak signal intensities of one of the two losses prohibited the correct assignment of the monostopic peak leading to false positive results. It should be mentioned that the mass of the precursor ion is deduced from the MS spectrum and not from the tandem mass spectrum. This is important because in many MS/MS experiments the precursor ion was entirely converted into fragment ions (Figure 5). PhosTShunter reported two peptides with masses of 2087.9240 and 2900.3345 Da leading to the identification of the



**Figure 6.** CAD spectra of phosphopeptides identified with our algorithm which were not identified with a database dependent search engine are shown. Panel A. The CAD spectra of the four times charged phosphopeptide EVVG[pS]AEAGVDAASVSEEF-RADHPFLF[camC]IK is identified by its  $y^{3+}$ -ion series. The phosphorylated serine can be localized within the sequence. The most prominent product ion is generated by the combined loss of water and phosphoric acid. Panel B. The spectrum of the five times charged peptide FDKLPGFGD[pS]IEAQ[camC]GTSVNVHSSLRDILNQITKPNDVYSFSLASR is shown. Panel C. The spectrum of the phosphopeptide FDKLPGFGD[pS]IEAQ[camC]GTSVNVHSSLRDILNQITKPNDVYSFSLASR with an additional modification at its cysteine residue is shown. The previously undescribed modification results in a mass shift of 151.977 Da most likely caused by covalent bound DTT.

phosphopeptides EVVG[pS]AEAGVDAASVSEEF ( $m/z$  1044.9705) and FDKLPGFGD[pS]IEAQ[camC]GTSVNVHSSLR ( $m/z$  726.0925) corresponding to two known phosphorylation sites S69 and S345 in ovalbumin. Additionally, the program identified the phosphopeptides EVVG[pS]AEAGVDAASVSEEF-RADHPFLF[camC]IK ( $m/z$  830.1415; Figure 6, Panel A) and FDKLPGFGD[pS]IEAQ[camC]GTSVNVHSSLRDILNQITKPNDVYSFSLASR ( $m/z$  1033.5145) with the molecular masses of 3316.5311 and 5162.5289 Da, respectively. These two peptides correspond to the same two phosphorylation sites but with an additional missed cleavage site. Interestingly, the program found the spectra of three additional phosphopeptides with the masses 3052.3358, 3468.5411, and 5314.5269 Da. Each of these three peptides was approximately 152 Da higher in mass than the three cysteine containing peptides described above. Analysis of the fragment ion spectra showed that these three peptides must have incorporated DTT in addition to carboxyamido-methylation of their cysteine residues (Figure 6, Panel B and Panel C). The theoretical mass shift of this modification is 151.9966 Da, fitting to the observed mean masses within 6 mDa. The y-ion series observed in these spectra allowed assignment of the mass shift corresponding to a covalently bound DTT to the cysteine residue (Figure 6, Panel C). In addition the program detected a phosphopeptide which a mass of 3620.5251 Da, pointing to the phosphopeptide EVVG[pS]AEAGVDAASVSEEF-RADHPFLFCIK covalently modified by two DTT moieties in addition to carboxyamido-methylation. The tandem mass spectrum of the peptide confirmed the primary structure and allowed the localization of this previously unknown modification to the cysteine residue.

In comparing the outcome of the Mascot search based on the same data, we report the following. Although Mascot identified the two smaller phosphopeptides with their phosphorylation sites correctly assigned to their sequence, the two larger phosphopeptides corresponding to the same phosphorylation sites were not identified. As a consequence, if the identified phosphorylation sites in these peptides had not provided only redundant information the respective phosphorylation sites would have been missed. More importantly, Mascot failed to identify the four peptides with the additional and unexpected modification at the cysteine residue. Although this modification is an artifact from the reduction and alkylation procedure and thus of no biological importance, the example nicely illustrates that PhosTShunter, being database independent, can identify phosphopeptides despite additional and even unknown modifications.

## Conclusions

The identification of phosphorylation sites in proteins is an extremely important but still demanding task in bioanalytical chemistry. The analytical problem is not only experimentally challenging but it includes a considerable task for data interpretation tools. Many of the state-of-the-art protocols in phosphoproteomics use LC-MS/MS in the last stage of the experimental routine. The large data files generated in these experiments have to be interpreted by automatic search engines. Obviously, these search engines will fail if the experimental data quality is poor or the additional modifications of a specific peptide are not known a priori or not included in the search parameters. The algorithm presented here is able to identify peptides phosphorylated at their serine or threonine residues in a database independent manner. Likewise, a priori assumptions regarding the primary structure of the peptides

are not required. The program uses data generated by a standard LC FT-ICR MS/MS experiment and applies two equally important criteria to CAD data, benefiting from the high resolution and mass accuracy of the FTMS data. PhosTShunter only reports detected phosphopeptides, leaving the task of the elucidation of the peptide sequence and the localization of the phosphorylated amino acids for a manual interpretation or de novo sequencing programs. Despite that, the algorithm is an extremely useful complementary tool to the existing search engines in PTM mapping due to its speed, specificity, sensitivity and reliability. It yields the number of detected phosphopeptides and their respective masses in the sample within a minute upon the completion of the standard LC-MS/MS run, thus providing an important input for decision making of the analyst and facilitates eventual manual interpretation of the data files.

**Acknowledgment.** We thank other members of our laboratories for help and fruitful discussions and Christopher Adams for critical reading of the manuscript. This work was supported by the Wallenberg Consortium North (Grant No. WCN2003-UU/SLU-009 to R. A. Z.). The purchase of the LTQ-FT instrument was supported by the Knut and Alice Wallenberg Foundation. T.K. was a recipient of a postdoctoral fellowship of the European Molecular Biology Organization (EMBO).

## References

- Johnson, S. A.; Hunter, T. Kinomics: methods for deciphering the kinome. *Nat. Methods* **2005**, *2*, 17–25.
- Blume-Jensen, P.; Hunter, T. Oncogenic kinase signaling. *Nature* **2001**, *411*, 355–365.
- Lee, H. G.; Perry, G.; Moreira, P. I.; Garrett, M. R.; Liu, Q.; Zhu, X.; Takeda, A.; Nunomura, A.; Smith, M. A. Tau phosphorylation in Alzheimer's disease: pathogen or protector? *Trends Mol. Med.* **2005**, *11*, 164–169.
- Hermiston, M. L.; Xu, Z.; Weiss, A. CD45: a critical regulator of signaling thresholds in immune cells. *Annu. Rev. Immunol.* **2003**, *21*, 107–137.
- Hunter, T. Signaling- -2000 and beyond. *Cell* **2000**, *100*, 113–127.
- Hutchins, J. R.; Clarke, P. R. Many fingers on the mitotic trigger: post-translational regulation of the Cdc25C phosphatase. *Cell Cycle* **2004**, *3*, 41–45.
- Rane, S. G.; Reddy, E. P. JAKs, STATs and Src kinases in hematopoiesis. *Oncogene* **2002**, *21*, 3334–3358.
- Plum, L.; Schubert, M.; Bruning, J. C. The role of insulin receptor signaling in the brain. *Trends Endocrinol. Metab.* **2005**, *16*, 59–65.
- Xie, Z.; Tsai, L. H. Cdk5 phosphorylation of FAK regulates centrosome-associated microtubules and neuronal migration. *Cell Cycle* **2004**, *3*, 108–110.
- Rane, S. G.; Reddy, E. P. Janus kinases: components of multiple signaling pathways. *Oncogene* **2000**, *19*, 5662–5679.
- Mann, M.; Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **2003**, *21*, 255–261.
- Cantin, G. T.; Yates, J. R. Strategies for shotgun identification of post-translational modifications by mass spectrometry. *J. Chromatogr. A* **2004**, *1053*, 7–14.
- Garcia, B. A.; Shabanowitz, J.; Hunt, D. F. Analysis of protein phosphorylation by mass spectrometry. *Methods* **2005**, *35*, 256–264.
- Jensen, O. N. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.* **2004**, *8*, 33–41.
- Steen, H.; Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell. Biol.* **2004**, *5*, 699–711.
- Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.
- Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
- Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.* **1998**, *120*, 3265–3266.
- Cooper, H. J.; Hakansson, K.; Marshall, A. G. The role of electron capture dissociation in biomolecular analysis. *Mass Spectrom. Rev.* **2005**, *24*, 201–222.
- Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and protein sequence analysis by electron-transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528–9533.
- Stensballe, A.; Jensen, O. N.; Olsen, J. V.; Haselmann, K. F.; Zubarev, R. A. Electron capture dissociation of singly and multiply phosphorylated peptides. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1793–1800.
- Mirgorodskaya, E.; Roepstorff, P.; Zubarev, R. A. Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. *Anal. Chem.* **1999**, *71*, 4431–4436.
- Hunter, A. P.; Games, D. E. Chromatographic and mass spectrometric methods for the identification of phosphorylation sites in phosphoproteins. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 559–570.
- Wilm, M.; Neubauer, G.; Mann, M. Parent ion scans of unseparated peptide mixtures. *Anal. Chem.* **1996**, *68*, 527–533.
- Neubauer, G.; Mann, M. Mapping of phosphorylation sites of gel-isolated proteins by nanoelectrospray tandem mass spectrometry: potentials and limitations. *Anal. Chem.* **1999**, *71*, 235–242.
- Steen, H.; Kuster, B.; Fernandez, M.; Pandey, A.; Mann, M. Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode. *Anal. Chem.* **2001**, *73*, 1440–1448.
- Borchers, C.; Parker, C. E.; Deterding, L. J.; Tomer, K. B. Preliminary comparison of precursor scans and liquid chromatography-tandem mass spectrometry on a hybrid quadrupole time-of-flight mass spectrometer. *J. Chromatogr. A* **1999**, *854*, 119–130.
- Tholey, A.; Reed, J.; Lehmann, W. D. Electrospray tandem mass spectrometric studies of phosphopeptides and phosphopeptide analogues. *J. Mass Spectrom.* **1999**, *34*, 117–123.
- Schlosser, A.; Pipkorn, R.; Bossemeyer, D.; Lehmann, W. D. Analysis of protein phosphorylation by a combination of elastase digestion and neutral loss tandem mass spectrometry. *Anal. Chem.* **2001**, *73*, 170–176.
- Bateman, R. H.; Carruthers, R.; Hoyes, J. B.; Jones, C.; Langridge, J. I.; Millar, A.; Vissers, J. P. A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 792–803.
- Kocher, T.; Allmaier, G.; Wilm, M. Nanoelectrospray-based detection and sequencing of substoichiometric amounts of phosphopeptides in complex mixtures. *J. Mass Spectrom.* **2003**, *38*, 131–137.
- Posewitz, M. C.; Tempst, P. Immobilized gallium(III) affinity chromatography of phosphopeptides. *Anal. Chem.* **1999**, *71*, 2883–2892.
- Goshe, M. B.; Conrads, T. P.; Panisko, E. A.; Angell, N. H.; Veenstra, T. D.; Smith, R. D. Phosphoprotein isotope-coded affinity tag approach for isolating and quantitating phosphopeptides in proteome-wide analyses. *Anal. Chem.* **2001**, *73*, 2578–2586.
- Kussmann, M.; Hauser, K.; Kissmehl, R.; Breed, J.; Plattner, H.; Roepstorff, P. Comparison of in vivo and in vitro phosphorylation of the exocytosis-sensitive protein PP63/parafusin by differential MALDI mass spectrometric peptide mapping. *Biochemistry* **1999**, *38*, 7780–7790.
- MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R. Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7900–7905.
- Ficarro, S. B.; McClelland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **2002**, *20*, 301–305.
- Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R. Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **1999**, *17*, 676–682.
- Opiteck, G. J.; Lewis, K. C.; Jorgenson, J. W.; Andereg, R. J. Comprehensive on-line LC/LC/MS of proteins. *Anal. Chem.* **1997**, *69*, 9, 1518–1524.

- (39) Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **1994**, *66*, 4390–4399.
- (40) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom. Rev.* **1998**, *17*, 1–35.
- (41) Nielsen, M. L.; Savitski, M. M.; Zubarev, R. A. Improving Protein Identification Using Complementary Fragmentation Techniques in Fourier Transform Mass Spectrometry. *Mol. Cell. Proteomics* **2005**, *4*, 835–845.
- (42) Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. 2005 New database-independent, sequence-tag-based scoring of peptide MS/MS data validates mowse scores, recovers below-threshold data, singles out modified peptides and assesses the quality of MS/MS techniques. *Mol. Cell. Proteomics* **2005**, *4*, 1180–1188.
- (43) Creasy, D. M.; Cottrell, J. S. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* **2002**, *2*, 1426–1434.
- (44) Annan, R. S.; Carr, S. A. Phosphopeptide analysis by matrix-assisted laser desorption time-of-flight mass spectrometry. *Anal. Chem.* **1996**, *68*, 3413–3421.
- (45) Chang, E. J.; Archambault, V.; McLachlin, D. T.; Krutchinsky, A. N.; Chait, B. T. Analysis of protein phosphorylation by hypothesis-driven multiple-stage mass spectrometry. *Anal. Chem.* **2004**, *76*, 4472–4483.
- (46) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10313–10317.

PR0503836