Confident Site Localization Using a Simulated Phosphopeptide Spectral Library

Suni, V.; Imanishi, S.Y.; Maiolica, A.; Aebersold, R.; Corthals, G.L.

DOI
10.1021/acs.jproteome.5b00050

Publication date
2015

Document Version
Final published version

Published in
Journal of Proteome Research

Journal of proteome research

Technical Note

pubs.acs.org/jpr

Confident Site Localization Using a Simulated Phosphopeptide Spectral Library

Veronika Suni,[†,‡,#] Susumu Y. Imanishi,[†,#] Alessio Maiolica,[§] Ruedi Aebersold,[§,∥] and Garry L. Corthals*[,†,⊥]

[†]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Tykistokatu 6, FI-20520 Turku, Finland

[‡]Turku Centre for Computer Science, Joukahaisenkatu 3-5 B, FI-20520 Turku, Finland

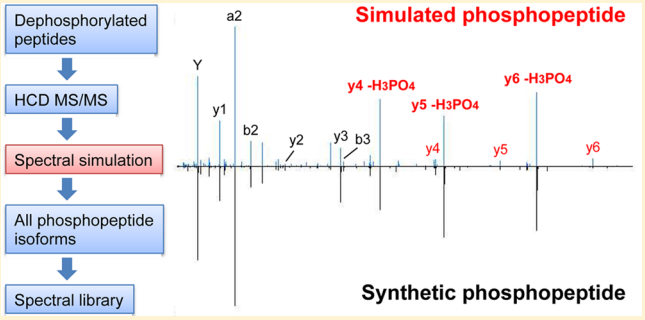
[§]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, HPT E 51, Wolfgang-Pauli-Strasse 16, 8093 Zurich, Switzerland

[∥]Faculty of Science, University of Zurich, 8057 Zurich, Switzerland

[⊥]Van 't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

Supporting Information

ABSTRACT: We have investigated if phosphopeptide identification and simultaneous site localization can be achieved by spectral library searching. This allows taking advantage of comparison of specific spectral features, which would lead to improved discrimination of differential localizations. For building a library, we propose a spectral simulation strategy where all possible single phosphorylations can be simply and accurately (re)constructed on enzymatically dephosphorylated peptides, by predicting the diagnostic fragmentation events produced in beam-type CID. To demonstrate the performance of our approach, enriched HeLa phosphopeptides were dephosphorylated with alkaline phosphatase and analyzed with higher energy collisional dissociation (HCD), which were then used for creating a spectral library of simulated phosphopeptides. Spectral library searching using SpectraST was performed on data sets of synthetic phosphopeptides and the HeLa phosphopeptides, and subsequently compared to Mascot and Sequest database searching followed by phosphoRS and Ascore afforded localization, respectively. Our approach successfully led to accurate localization, and it outperformed other methods, when phosphopeptides were covered by the library. These results suggest that the searching with simulated spectral libraries serves as a crucial approach for both supplementing and validating the phosphorylation sites obtained by database searching and localization tools. For future development, simulation of multiply phosphorylated peptides remains to be implemented.



KEYWORDS: *beam-type CID, LC−MS/MS, phosphoproteomics, phosphorylation site localization, spectral library searching, spectral simulation*

INTRODUCTION

For two decades now sequence database searching has been the method of choice for semiautomated peptide and protein identification in tandem mass spectrometry (MS/MS) based proteomics (Figure 1A).[1−3] Recently, a new type of search algorithms based on spectral library matching has become available.[4−6] Unlike sequence database search engines that compare MS/MS spectra to theoretical information based on a sequence database, spectral library methods match experimental MS/MS spectra to a library of MS/MS spectra derived from previous identifications. By comparing two experimental data, the spectral library matching approach takes advantage of comparison of specific spectral features that are often not used or neglected in database searches, including actual peak intensities, neutral losses from fragments, and various uncommon or even unknown fragments characteristic to specific peptides, thus improving and expediting the identification of the best match.[7] The method has also been shown to be effective for the analysis of phosphopeptides, as exemplified in PhosphoPep[8,9] and ProMEX[10] (Figure 1A), because the similarity scores of spectral searching algorithms are more precise than classical database search approaches. Nevertheless, while these phosphopeptide databases are extensive and some can be downloaded (e.g., PhosphoPep from PeptideAtlas[11]) and then searched by SpectraST,[4] all phosphopeptide libraries at this time need to be considered incomplete and are not yet cross platform compatible.

Received: January 20, 2015
Published: March 16, 2015

DOI: 10.1021/acs.jproteome.5b00050
J. Proteome Res. 2015, 14, 2348−2359

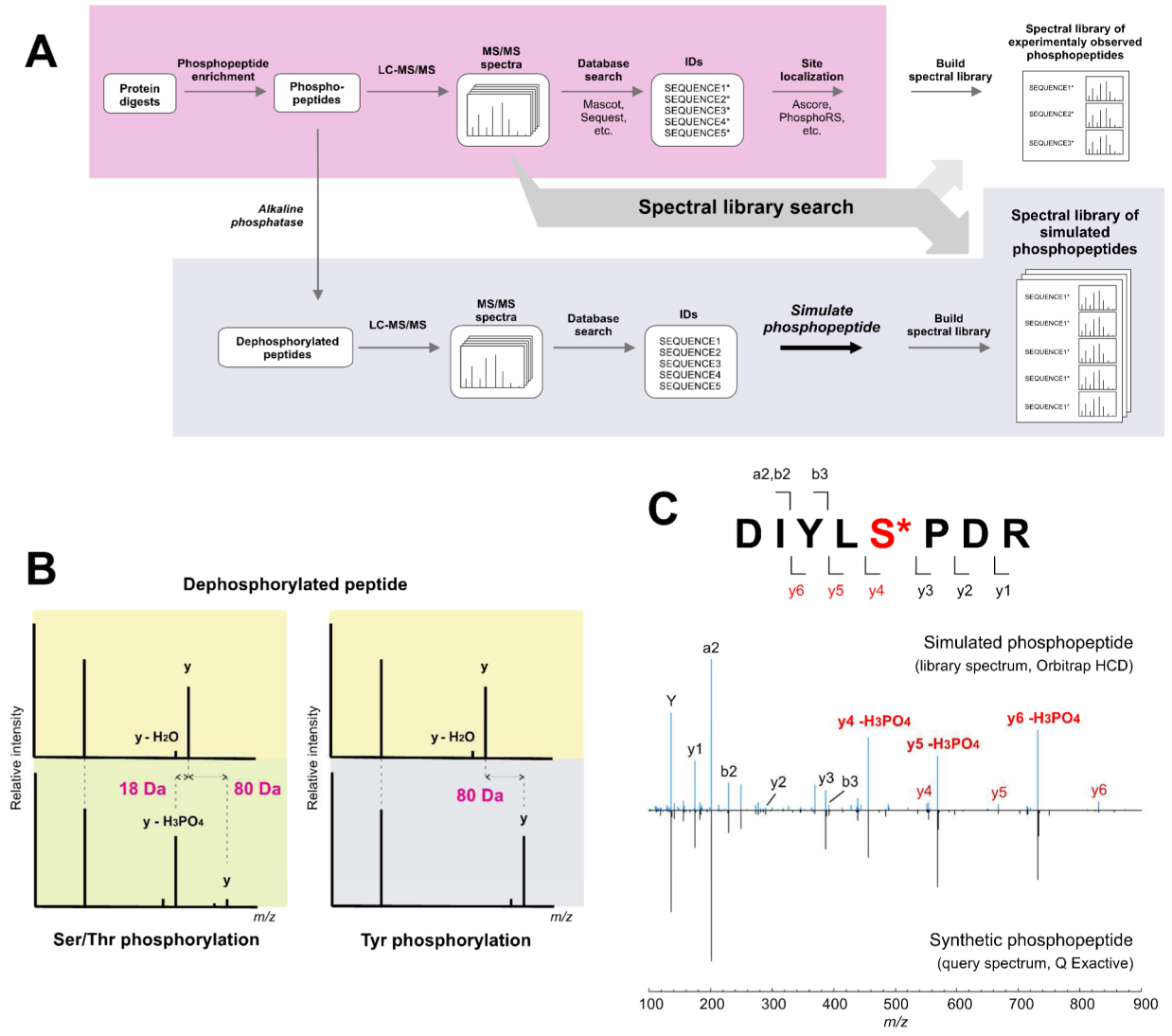


Figure 1. Simulated spectral library for site-specific identification of phosphopeptides. (A) A typical workflow for MS/MS-based phosphorylation analysis includes sequence database searching and site localization of enriched tryptic phosphopeptides (upper). In contrast, our strategy is based on spectral library searching, in which confident and sensitive localization is enabled by spectral simulation of all possible single phosphorylations (lower). (B) The simulation of single phosphorylation is performed on HCD spectra of enzymatically dephosphorylated peptides because of the high spectral similarity and identification efficiency (Supporting Information Figures 2 and 3). We predict the position of phosphorylated fragment ions by a mass shift of +80 Da (e.g., +40 m/z when doubly charged) with 10% (phospho Ser/Thr) or 100% (phospho Tyr) intensity. A neutral loss of phosphoric acid (−98 Da) is predicted by a shift of −18 Da with the maintained intensity (phospho Ser/Thr). (C) A representative HCD spectral match of a simulated HeLa phosphopeptide (DIYLS*PDR; S*: phosphoserine) with its synthetic phosphopeptide represents well predicted fragment ions in the simulated spectrum. SimSpectraST searching of the 20 synthetic phosphopeptide HCD data was performed with SimHeLa library (refer to Tables 1 and 2). See other representative spectral matches in Supporting Information Figure 5.

To achieve high confidence spectral searching, a library of high-quality MS/MS spectra with accurately identified peptide data must be created. This can be time-consuming and costly, particularly for phosphopeptide data, as the precise localization of phosphorylation sites remains a difficult task even for most but a few expert mass spectrometrists. Recently, a concept for creating a spectral library of computationally simulated phosphopeptides has been reported by Hu and Lam.[12,13] For the simulation of known phosphorylation sites, they used MS/MS spectra of nonphosphorylated peptides generated by ion trap collision-induced dissociation (CID), in which masses of fragment ions were shifted by +80 Da to predict fragment ions retaining a phosphate group. SpectraST searching against their spectral library supplemented with simulated spectra was proven to be highly sensitive for the identification of phosphopeptides compared to database searching.[13] When this approach (called semiempirical prediction) was applied to simulation of unmodified peptides with amino acid substitutions, it also outperformed computational modeling of CID spectra using Mass Analyzer,[14,15] in spectral similarity and identification sensitivity.[12] However, the accuracy of the phosphorylation simulation and the reliability in the phosphorylation site localization have not been investigated yet.

Previously we have demonstrated that similar, but not identical, fragmentation patterns are observed when phosphopeptides and their enzymatically dephosphorylated counterparts (i.e., nonphosphorylated peptides) are compared. Specifically, these observations include the presence of fragment ions in spectra acquired in a quadrupole time-of-flight (Q-TOF) instrument, and their intensity distribution.[16]

DOI: 10.1021/acs.jproteome.5b00050
J. Proteome Res. 2015, 14, 2348−2359

Fragment ions containing the phosphorylated serine and threonine residues show a loss of 18 Da in comparison to the corresponding fragment ions derived from the dephosphorylated peptide (Figure 1B). This loss occurs with maintained intensity because, along with a prominent neutral loss of phosphoric acid from those residues,[17] a beam-type quadrupole collision cell of the Q-TOF instrument induces accompanying fragmentation of the peptide backbone; therefore, the loss from fragment ions can be used to confidently localize the phosphorylation sites (reference-facilitated phosphoproteomics).[16]

Here, we propose a strategy for simple, but accurate, simulation of phosphopeptides by resembling the diagnostic fragmentation events in beam-type CID (Figure 1A,B). In this strategy, all possible single phosphorylations are computationally simulated on dephosphorylated peptides. By matching to accurately simulated spectra, SpectraST is enabled to localize phosphorylation sites on actual phosphopeptides (Figure 1C). To demonstrate the performance of this method, we have obtained dephosphorylated peptides derived from HeLa cells and acquired beam-type CID spectra for those peptides, which were then used for creating a spectral library after the phosphorylation simulation. SpectraST searching using the simulated spectral library was tested on data sets of synthetic phosphopeptides and HeLa phosphopeptides, and evaluated in comparison to other existing methods including sequence database searching followed by phosphorylation site localization.

EXPERIMENTAL SECTION

Sample Preparation

HeLa cells were grown in DMEM (GIBCO) supplemented with 10% fetal calf serum (Sigma), at 37 °C in 5% CO_2. Cells were washed twice with ice-cold phosphate buffered saline (PBS) and lysed directly on 15 cm plates with 400 μL of buffer containing 100 mM ammonium bicarbonate (AMBIC), pH 8, 5 mM EDTA, a phosphatase inhibitor mix (Roche) and 8 M urea. The cells were scraped from the dish surface and the solution transferred in Eppendorf tube and sonicated with 3 cycles of 30 s each on ice. The protein extract was centrifuged at 16 000*g* for 30 min. The supernatant was collected and protein concentration measured with BCA assay (Invitrogen). The proteins were reduced by incubation with tris(2-carboxyethyl)-phosphine (TCEP) at a final concentration of 10 mM at room temperature for 45 min. The produced free thiols were alkylated with 20 mM iodoacetamide at room temperature for 45 min in the dark. The solution was diluted with 50 mM AMBIC pH 8, to a final concentration of 1 M urea and digested with sequencing grade-modified trypsin (Promega) overnight at 37 °C. Peptides were desalted on a C18 Sep-Pak cartridge (Waters) and dried in a Speedvac. For phosphopeptide enrichment experiment, 10 mg of total peptide was used. Dried peptides were reconstituted in a solution consisting of 80% acetonitrile (ACN), 3.5% TFA saturated with phthalic acid. About 3 mg of TiO_2 resin (GL Sciences) was placed into a 1.5 mL Eppendorf tube and was subsequently washed with 300 μL of 2-propanol and equilibrated with 280 μL of the saturated phthalic acid solution for 30 min. The resin was successively transferred to the peptide solution and incubated on rotation for 45 min at room temperature. After incubation step the resin was centrifuged at 100*g* for 1 min and the peptide solution was discarded. The resin was subsequently thoroughly washed two times each with 280 μL of the phthalic acid solution and two times with 80% ACN, 0.1% TFA solution, and finally two times 50% ACN, 0.1% TFA. In the final step, phosphopeptides were eluted from the TiO_2 resin using two times 150 μL of a 0.3-M NH_4OH solution (pH 10.5). After elution, the pH of the pooled eluates was rapidly adjusted to 2.7 using 50% TFA, and phosphopeptides were purified using an appropriate reverse phase column suitable for up to 100 μg of total peptide (C18Micro Spin Column, Nest Group). The peptides were desalted as before but in addition that the last two column washes were performed with water and the peptides eluted with 300 μL of 80% ACN in water, without any acid to consent to easily adjust the pH to 7.9 for successive enzymatic reaction. The eluted sample was divided in two and dried in SpeadVac. To dephosphorylate peptides, one of those was reconstituted in 150 μL of 100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM dithiothreitol pH 7.9 and incubated with 5 μL of alkaline phosphatase (New England Biolabs) for 3 h at 37 °C. The solution was successively acidified using 50% TFA and again desalted. The dried peptides were reconstituted in 2% ACN, 0.5% formic acid (FA) and analyzed by LC−MS/MS.

Mass Spectrometry

LC−MS/MS analyses for the HeLa phosphopeptide and dephosphorylated peptide samples were performed in duplicate by submitting 1 μL of the peptides (estimated concentration 1−2 μg/μL). An EASY-nLC II nanoflow LC instrument was coupled to an electrospray ionization (ESI) mass spectrometer LTQ Orbitrap Velos (Thermo Fisher Scientific). A 100 μm × 3 cm trap column and a 75 μm × 15 cm analytical column were in-house packed with Magic C18AQ resin (200 Å, 5 μm; Michrom Bioresources). The mobile phases were 2% ACN, 0.2% FA (A) and 95% ACN, 0.2% FA (B). LC gradient elution condition was initially 2% B to 20% B (70 min), 40% B (100 min), and 100% B (105−110 min), with a flow rate of 300 nL/min. Data dependent acquisition was performed in positive ion mode. MS spectra were acquired from m/z 300 to m/z 2000 in the Orbitrap set to a resolution of 30 000 at m/z 400 with a target value of 1 000 000 ions and a maximal injection time of 100 ms in profile mode. The 10 most abundant ions of which charge states were 2+ or higher were selected for subsequent fragmentation, and MS/MS spectra were acquired in the Orbitrap with a resolution of 7500 at m/z 400, a target value of 50 000 ions, a maximal injection time of 200 ms, and the lowest mass fixed at m/z 100, in centroid mode. Four fragmentation techniques were used; CID with normalized collision energy of 35 and activation time of 10 ms, multistage activation (MSA) with normalized collision energy of 35 and activation time of 10 ms for the precursors and precursor −98/z,[18] higher energy collisional dissociation (HCD) with normalized collision energy of 40 and activation time of 0.1 ms, and electron-transfer dissociation (ETD) with supplemental activation and activation time of 100 ms (charge state dependent). Note, the instrument control software LTQ Tune Plus version 2.6.0.1050 was used in this study, and the HCD energy of 40 on that version would be equivalent to 35 on updated versions, e.g., 2.6.0.1065 SP3. Dynamic exclusion duration was 60 s. The lock-mass option was used (m/z 445.120025).

For analysis of synthetic phosphopeptides (PEPotec, Thermo Fisher Scientific), LC−MS/MS was performed using the Orbitrap system and also an EASY-nLC 1000 nanoflow LC instrument coupled to a Q Exactive ESI-quadrupole-orbitrap mass spectrometer (Thermo Fisher Scientific). Twenty

DOI: 10.1021/acs.jproteome.5b00050
J. Proteome Res. 2015, 14, 2348−2359

Table 1. LC−MS/MS Data Sets Used in This Study

data set	fragmentation	description of sample	reference
		HeLa	
HeLa phosphopeptide data sets	HCD, ETD, CID, MSA	Phosphopeptides were enriched from a HeLa tryptic digest	
HeLa dephosphorylated peptide data sets	HCD, ETD, CID, MSA	Enriched HeLa phosphopeptides were dephosphorylated	
		Synthetic	
20 synthetic phosphopeptide data sets	HCD, CID	20 singly phosphorylated peptides were selected and synthesized	Supporting Information Table 1
Marx data set[a]	HCD	>100 000 of singly phosphorylated peptides and nonphosphorylated counterparts were synthesized	Marx et al., 2013 (ref25)

[a]Downloaded from the PRIDE database.

synthetic singly phosphorylated peptides were mixed into 2 groups (10 peptides each), where phosphopeptide isoforms were separated from each other (see Supporting Information Table 1 for the peptide selection). Both the mixtures (1 pmol peptide each, based on the concentration given by the manufacturer) were analyzed in triplicates. The columns and the mobile phases mentioned above were used. LC gradient elution condition was initially 5% B to 35% B (10 min), and 100% B (12−20 min), with a flow rate of 300 nL/min. Data dependent acquisition was performed as described above, with some modifications. Dynamic exclusion was disabled. In the Orbitrap system, MS spectra were acquired in the Orbitrap, the 20 most abundant ions were selected for subsequent CID, and then MS/MS spectra were acquired in the LTQ ion trap with a target value of 5000 ions and a maximal injection time of 50 ms, in centroid mode. In the Q Exactive system, MS spectra were acquired from m/z 300 to m/z 2000 with a resolution of 70 000 at m/z 200, a target value of 1 000 000 ions, and a maximal injection time of 120 ms, in profile mode. The 10 most abundant ions were selected for subsequent fragmentation (HCD) with normalized collision energy of 30. MS/MS spectra were acquired with a resolution of 17 500 at m/z 200, a target value of 50 000 ions, a maximal injection time of 250 ms, and the lowest mass fixed at m/z 100, in profile mode. All LC−MS/MS data sets used in this study are listed in Table 1.

Database Search

Orbitrap and Q Exactive data files were converted to mzXML files using Trans Proteomic Pipeline (TPP)[19] version 4.4 VUVUZELA revision 1, and then to mgf files using MzXML2Search, with the default parameters, except for centroiding the fragment ions acquired by Q Exactive in profile mode. The HeLa phosphopeptide and dephosphorylated peptide data acquired with the four fragmentation methods (see Table 1) were searched with Mascot v2.4.1 (Matrix Science) against a concatenated forward−reverse SwissProt database (release date March 2010, *Homo sapiens*, total 40530 sequences). But, search results of the dephosphorylated peptide HCD data used for creating a simulated HeLa spectral library (see below) were obtained also with an older version of Mascot (v2.2.6) before the software update. The following Mascot search condition was used for both the phosphopeptide and dephosphorylated peptide data: carbamidomethylation (C) as a fixed modification, and oxidation (M), phosphorylation (S, T, and Y) and acetylation (protein N-terminus) as variable modifications. Trypsin was specified as an enzyme and two missed cleavage sites were allowed. Peptide mass tolerance and MS/MS ion tolerance were set to 50 ppm and 0.05 Da, respectively. Sequest searches of the HeLa phosphopeptide HCD data were done using the SORCERER platform v4.2.0 (Sage-N Research) with corresponding search parameters. To reduce a bias at the identification step prior to the investigation of phosphorylation site localization, the same identification condition was applied to different programs, i.e., a false discovery rate (FDR) of 1% was estimated using the target-decoy strategy by PeptideProphet.[20] For each peptide spectral match, a discriminant score is calculated which is in turn used to assign probability values. Minimum probabilities corresponding to 1% FDR were applied as cutoffs, e.g., 0.80 and 0.90 for the Mascot and Sequest search results of the HeLa phosphopeptide HCD data, respectively,. For phosphorylation site localization, Ascore[21] v2008.04.16f BETA (Sage-N Research) was used on the Sequest results. Another localizaton tool phosphoRS[22] v3.0 was used on the Mascot search results, by submitting the mgf files of the HeLa phosphopeptide HCD data to Proteome Discoverer v1.4.0.288 (Thermo Fisher Scientific) after adjusting mgf file formats. The phosphoRS option for considering H_3PO_4 neutral loss ions was disabled (refer to the Results and Discussion for the evaluation of that option). Also, to comparatively evaluate Mascot delta scores of nonphosphorylated peptides with those obtained from a large synthetic peptide data set (see below), Mascot searches of the HeLa dephosphorylated peptide HCD data were repeated via Proteome Discoverer v1.4.

For the 20 synthetic phosphopeptide HCD data obtained using the Q Exactive (see Table 1), Ascore v2011.01.12a BETA was used on Sequest in the same condition. PhosphoRS (the initial version implemented in the Proteome Discoverer v1.3.0.339) and phosphoRS v3.0 (the neutral loss option enabled or disabled) were used on Mascot v2.4.0 in the same condition, except for submitting the Q Exactive raw data to Proteome Discoverer without the file conversion. Also, to evaluate PTM score,[23] MaxQuant[24] v1.3.0.5 (Andromeda searching) was used by submitting the raw data with corresponding search parameters, but the data were deconvoluted before extraction of the ten most abundant peaks per 100 Th as described by Marx et al.[25] Since the sequences and phosphorylation sites are known and the data complexity was low, all statistical filters were disabled. In this study, a large and complex synthetic peptide data set was also used, which required a different database search condition (described below).

Phosphorylation Simulation Algorithm

HCD MS/MS spectra of the HeLa dephosphorylated peptides identified by Mascot were modified to resemble spectra of singly phosphorylated peptides (Figure 1B). For that, peaks of main ions (a, b, y, and remaining precursor ions; neutral loss of water; monoisotopic and 3 isotopic ions; singly and multiply charged up to precursor charge but not more than 3) were

DOI: 10.1021/acs.jproteome.5b00050
J. Proteome Res. 2015, 14, 2348−2359

Table 2. Spectral Libraries Used in This Study

spectral library	number of spectra	description of source data	reference
		Experimental	
Synthetic phosphopeptide spectral library	31	HCD spectra of 20 synthetic phosphopeptides	Supporting Information Table 1
		Simulated	
SimHela library	23 126[a]	Phosphopeptide HCD spectra simulated based on HeLa dephosphorylated peptides	
SimMarx library	285 252[a]	Phosphopeptide HCD spectra simulated based on synthetic nonphosphorylated peptides of Marx data set	Marx et al., 2013 (ref25)
		Combined	
SimHeLa-mouse-yeast library	162 789[a]	SimHeLa library merged with CID spectral libraries of mouse and yeast phosphopeptides[b]	Hu and Lam, 2013 (ref13)
Hu&Lam library	106 330[a]	Libraries of experimentally obtained and simulated CID spectra of phosphopeptides[b]	Hu and Lam, 2013 (ref13)

[a]Decoy entries included. [b]Downloaded from the PeptideAtlas database.

shifted by −18.0106 Da according to the site on which phosphorylation was being simulated. The shifted peaks correspond to ions with a neutral loss of phosphoric acid ($-H_3PO_4$, −97.9769 Da) that are expected to be present in MS/MS spectra of serine and threonine phosphorylated peptides. The main ions maintaining a phosphate group ($+HPO_3$) were formed by creating peaks with masses 79.9663 Da greater than the ions in dephosphorylated peptide with 10% of their original intensity. When tyrosine phosphorylation was simulated, a different shift was performed resulting in the main ions maintaining a phosphate group ($+HPO_3$) but not the neutral loss. In addition, the tyrosine immonium ion was shifted as well.

The peaks to be shifted were searched by their theoretical masses in the peak lists of the original MS/MS spectra within a mass window of ±0.05 m/z. Fragment ion m/z values were calculated according to Roepstorff−Fohlmann−Biemann nomenclature.[26,27] A modification at residue i means that there is a mass shift in the b series of ions b_i to b_{n-1}, a series of ions a_i to a_{n-1} and in the y series of y_{n-i+1} to y_{n-1}, where n is the number of residues in the peptide (see Figure 1B). The precursor mass increase by 79.9663 Da was taken into account.

The algorithm was implemented using Proteowizard,[28] which is an external set of TPP software libraries and tools. The code of the converter **msconvert** was modified to add simulation functionality. The program **SimPhospho** and an Excel macro (for calculating deltaDot score, described below) will be available from our Web site (http://www.btk.fi/proteomics/) and a source code can be obtained from the authors upon request. The detailed workflow is summarized in Supporting Information Figure 1.

Spectral Library of Simulated HeLa Phosphopeptides

Spectral libraries of the simulated HeLa phosphopeptides were built using functionalities of SpectraST (version 4.0, released beta). The detailed workflow is summarized in Supporting Information Figure 1. Raw spectral libraries were generated from the two replicate HCD runs of the dephosphorylated HeLa sample, after applying the phosphorylation simulation to PeptideProphet validated Mascot search results with minimum probability of 0.90. These raw libraries were joined together to form a library, where spectral replicates of the same peptide ion were combined into a representative consensus spectrum.[7] Since peptides having protein N-terminal acetylation were not recognized by the SpectraST used in this study, those peptides were not taken into account for the library (also in all search results obtained in this study). Finally, to allow FDR estimation of spectral search results, decoy entries were added to the consensus library using built-in function on SpectraST[29] in the following way. Each peptide sequence is shuffled randomly to form a decoy sequence. After that the fragment ion peaks are repositioned according to that decoy sequence. The resulting spectral library (SimHeLa library) consisted of total 23 126 spectra: 11 563 consensus spectra of simulated phosphopeptides (4729 and 5410 spectra of doubly and triply charged phosphopeptides, respectively, and 567 spectra of phosphopeptides with methionine oxidation) and the other half 11 563 was artificially generated spectra of decoy entries. Only 2495 spectra (22% of the total) originated from single spectral replicates, 8327 (72%) had 2−3 replicates, 699 (6%) had 4−9 replicates, 33 had 10−19 replicates, and 9 spectra had 20 or more replicates based on which the consensus spectra were built. Also, to comparatively evaluate SimHeLa library with and without adding large sets of spectra, ion trap CID spectral libraries of mouse and yeast phosphopeptides downloaded from PeptideAtlas (51 420 and 18 412 spectra, respectively; release date 2013.07.15)[13] were supplemented with decoy entries and then joined to SimHeLa library using SpectraST (SimHeLa-mouse-yeast library). All spectral libraries used in this study are listed in Table 2.

Spectral Library Search for HeLa Data

The HeLa phosphopeptide HCD data (mzXML) were searched against SimHeLa and SimHeLa-mouse-yeast libraries by SpectraST with the default parameters. The results were validated by PeptideProphet, where an FDR of 1% required minimum probability of 0.70 and 0.80 for the results obtained with SimHeLa and SimHeLa-mouse-yeast libraries, respectively. A customized deltaDot score cutoff was applied to the spectral search results. DeltaDot score contributes to the total discriminant score (F-value) for every spectral match[30] and reflects how much the first matching peptide spectrum differs from the second best match in dot product scores. When the deltaDot score is equal to 0, it means that those spectral matches are not discriminated due to the identical dot score. DeltaDot score used in SpectraST considers a match to a nonhomologous peptide as the second best hit, therefore it is not applicable for assessing a difference in the quality of a match of phosphopeptides with alternative phosphorylation sites. Due to that, we recalculated deltaDot score by comparing first and second best hits even if they represent the same peptide sequence with only difference in positions of phosphorylation sites. Occasionally, due to a lack of the second best hits in SpectraST search results exported with the default

DOI: 10.1021/acs.jproteome.5b00050
J. Proteome Res. 2015, 14, 2348−2359

parameters, the recalculated score would be higher than the original. In this case, we used the original deltaDot score instead of the recalculated. A recalculated deltaDot score was selected as a prerequisite cutoff used for phosphorylation site localization. The detail of the whole workflow is presented in Supporting Information Figure 1. For sharing viewable annotated spectral matches, the SpectraST search of the HeLa phosphopeptide HCD data was repeated using SimHeLa library by submitting the mgf files to Proteome Discoverer 1.4, and the search result (an msf file) was deposited to the ProteomeXchange (see below). This file covered all the spectral matches at 1% FDR and 1% FLR (refer to the Results and Discussion), but there was difference in the scores.

Spectral Library Search for 20 Synthetic Phosphopeptide Data

For spectral searching of the 20 synthetic phosphopeptides, two spectral libraries were created in addition to the above-mentioned simulated spectral libraries (refer to Table 2). First, a spectral library was created based on the 20 synthetic phosphopeptide HCD data itself. The Q Exactive raw data were converted to mzXML and then mgf files, followed by searching with Mascot v2.4.1 against the SwissProt database as described above. All spectral replicates correctly assigned by Mascot were used for generating consensus spectra by SpectraST. However, cross-contamination between the two mixtures of phosphopeptide isoforms was suspected due to repeated false localization by Mascot search at certain retention time periods. A few percent of the contamination based on precursor ion intensity was confirmed by manually inspecting extracted ion chromatograms and MS/MS spectra on Xcalibur v2.2 (Thermo Fisher Scientific). Therefore, those spectral replicates were excluded (see Supporting Information Table 1). The resulting spectral library consisted of 31 consensus spectra, including the peptides with different charge and oxidation states, but no decoy was supplemented (Table 2). Also, ion trap CID spectral libraries of experimentally obtained and computationally simulated human phosphopeptides (18 066 and 35 099 spectra, respectively; release date 2013.07.15; developed by Hu and Lam[13]) were downloaded from PeptideAtlas. These downloaded libraries were joined together and then supplemented with decoy entries using SpectraST (Hu&Lam library, see Table 2).

In SpectraST searching of the 20 synthetic phosphopeptides, the HCD data (mzXML) were searched using four spectral libraries (synthetic phosphopeptide, Hu&Lam, SimHeLa, and SimHeLa-mouse-yeast libraries). Also, the CID data of the 20 synthetic phosphopeptides (converted to mzXML as well) were searched using two spectral libraries (Hu&Lam and SimHeLa libraries). No FDR cutoff was applied because of the known peptide sequences and data simplicity. The cross-contamination of phosphopeptide isoforms described above was manually excluded from the search results, since it has significant influence at low false localization rates (FLRs). FLR was calculated as follows: FL/(TL + FL) where TL and FL are the numbers of true and false localization spectra at a score threshold, respectively (spectra under the score threshold, i.e., ambiguous localization spectra, are not taken into consideration).

Spectral Library and Database Searches for a Large Synthetic Peptide Data Set

A large synthetic peptide data set generated by Marx et al.[25] was also used for evaluating the performance of SpectraST and phosphoRS. The 96 Orbitrap Velos HCD raw data were downloaded from PRIDE, which contain both the singly phosphorylated peptides and the nonphosphorylated counterparts (Marx data set, see also Table 1). A simulated phosphopeptide spectral library was created as described above, with some modifications. The raw data were converted to mzXML and then mgf files. The data were searched with Mascot v2.4.1, against the provided sequence database[25] containing concatenated synthetic peptide sequences as artificial proteins and human IPI protein sequences as decoy entries. The following Mascot search condition was used: oxidation (M) and phosphorylation (S, T, and Y) as variable modifications, maximum four missed cleavages by trypsin, and MS and MS/MS mass tolerances of 50 ppm and 0.05 Da, respectively. The mass tolerance was significantly increased from the condition reported by Marx et al. (5 ppm), since we found that there seemed to be mass calibration fluctuation (+10−20 ppm) in those data apparently due to the lock-mass calibration failure (this was also seen in our Orbtrap Velos data). PeptideProphet validated nonphosphorylated peptides with minimum probability of 0.95 were used for the phosphorylation simulation, after applying Mascot delta score of 10 as an additional cutoff. On the basis of 29 944 peptide sequences, a consensus spectral library of the simulated phosphopeptides was created with decoy entries (SimMarx library, total 285 252 spectra) and used for SpectraST searching as described above. The results were filtered based on the following Mascot/phosphoRS search result, where only phosphopeptide spectral matches sharing the same sequences between two searches were taken into consideration (also refer to Results and Discussion section).

The Mascot search of the Marx data set was repeated for evaluating phosphoRS v3 with and without the neutral loss considering, via Proteome Discoverer v1.4 after adjusting mgf file formats. An FDR of 1% was estimated using the target-decoy strategy separately for phosphorylated and nonphosphorylated peptides, as described by Marx et al.[25] The phosphopeptides identified with the synthetic peptide sequences were used for the phosphoRS evaluation. The nonphosphorylated peptides were used for evaluating Mascot delta scores at a peptide sequence level.

Data Sets Availability

Raw mass spectrometry data, protein sequence database, spectral libraries, and search results have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the data set identifier PXD000474. For viewing the annotated spectra in the msf result file, a free Proteome Discoverer viewer is available from the Thermo Omics Software Portal (http://portal.thermo-brims.com)

RESULTS AND DISCUSSION

LC−MS/MS Analysis of Phosphorylated and Dephosphorylated Peptides Derived from HeLa Cells

The experimental protocol used here is similar to the one described previously[16] (Figure 1A). Digested human HeLa cell proteins were enriched for phosphopeptides using TiO_2, and half of the sample was dephosphorylated by calf intestinal alkaline phosphatase treatment. Both samples were analyzed by LC−MS/MS (Table 1). An LTQ Orbitrap Velos instrument equipped with an octapole collision cell was used for acquiring MS/MS spectra generated by HCD (beam-type CID). While the Orbitrap has several fragmentation options including ion

DOI: 10.1021/acs.jproteome.5b00050
J. Proteome Res. 2015, 14, 2348−2359

trap CID, MSA (also ion trap CID), and ETD, the most striking similarity of phosphorylated and dephosphorylated peptides could be seen in HCD spectra, and also in ETD spectra (see representative spectra in Supporting Information Figure 2). Mascot database searches revealed that, at an FDR of 1% estimated by PeptideProphet, HCD outperformed all other fragmentation methods in the number of identifications, both in the analysis of phosphorylated and dephosphorylated peptides (Supporting Information Figure 3), a result that agrees with previous reports.[25,31] Given the high spectral similarity and the superior number of identifications, HCD was expected to be the most suitable fragmentation method for the spectral simulation of phosphorylation (refer to the Introduction and Figure 1 for the simulation strategy). The majority of phosphopeptides identified by HCD were singly phosphorylated (84%) (Supporting Information Figure 4). Eighty-six percent of the singly phosphorylated peptides had more than one possible phosphorylation site (S, T or Y), which require validation of the precise phosphorylation site location. With high efficiency dephosphorylation *in vitro* (99%), we observed HCD spectra for 81% of the singly phosphorylated peptides in their dephosphorylated form (Supporting Information Figure 4). This provided a data set that allowed for development of the phosphorylation simulation approach and for evaluation of its performance in spectral library searching.

Creating a Spectral Library of Simulated Phosphopeptides

On the basis of the HCD spectra obtained from the HeLa dephosphorylated peptides (Table 1), spectra of singly phosphorylated peptides were computationally simulated. The principle of the simulation is illustrated in Figure 1B and further details are described in Experimental Section. For building a simulated spectral library, stringent identification criteria were used to ensure high quality HCD spectra of the HeLa dephosphorylated peptides (PeptideProphet minimum probability of 0.90 on Mascot search results), resulting in identification of 2654 unique peptides at 0.35% FDR. On the basis of these HCD spectra, 11 563 consensus spectra simulated for singly phosphorylated peptides were generated to cover all the possible phosphorylation sites on the peptide sequences. The simulated spectra and 11 563 decoy entries were used to compile a spectral library (SimHeLa library, see Table 2).

To evaluate SimHeLa library, 20 singly phosphorylated peptides were synthesized and HCD spectra were acquired using a Q Exactive instrument (Table 1). The synthetic peptides were selected from differentially localized/identified phosphopeptides in preliminary analysis of the HeLa data set to clearly observe the difference in the performance of tested programs (see details in Supporting Information Table 1). LC−MS/MS analysis was performed in triplicates without using the dynamic exclusion function, resulting in acquisition of many spectral replicates for each phosphopeptide (>2000 spectra for 20 phosphopeptides, refer to Supporting Information Table 2). SpectraST searching of these synthetic phosphopeptide HCD data was performed with SimHeLa library. Representative spectral matches demonstrated well-predicted fragment ions in simulated spectra (Figure 1C and Supporting Information Figure 5). To investigate the spectral similarity, SimHeLa library was compared to a spectral library built from the same synthetic phosphopeptides (Table 2), and consequently SimHeLa library gave somewhat lower dot product scores in the searches (17% lower, see Supporting Information Figure 6A). Since phosphorylation sites on the synthetic phosphopeptides were known, performance in site localization could be investigated. At 1% FLR, SimHeLa library showed 17% lower sensitivity than the synthetic phosphopeptide spectral library did (Supporting Information Figure 6B). In this data analysis, 2 of the 20 phosphopeptides, which lacked alternative phosphorylation sites in the synthetic phosphopeptide spectral library (see Supporting Information Table 1), were excluded since the FLR calculation requires observation of false localization matches.

Furthermore, SimHeLa library was compared with the library previously reported by Hu and Lam, which consisted of experimentally obtained phosphopeptide spectra and those simulated for known phosphorylation sites based on non-phosphorylated peptide spectra (Hu&Lam library, see Table 2).[13] As Hu&Lam library was developed based on ion trap CID spectra, the 20 synthetic phosphopeptides were analyzed also by ion trap CID without using the dynamic exclusion function (Table 1). Out of the 20 phosphopeptides, 6 were excluded from data analysis as their sequences were not covered by Hu&Lam library. In SpectraST searching of these limited data sets, SimHeLa library outperformed Hu&Lam library, of which the localization sensitivity at 1% FLR was only 2.4% (on the CID data) as opposed to the method reported here which was 61% (on the HCD data) (Supporting Information Figure 7). This localization failure of Hu&Lam library was presumably due to the combination of false localization on experimental spectra stored in the library, inaccurate simulation on CID spectra (see representative spectral matches in Supporting Information Figure 8), and also incomplete coverage of phosphorylation sites on stored peptides. Due to the high spectral similarity of phosphopeptide isoforms (i.e., the same sequence with different phosphorylation sites), a missing phosphorylation site in a spectral library may induce false localization to the alternative sites stored in the library.

Taken together, these results suggest that HCD spectra of dephosphorylated peptides served as a basis for the accurate simulation of phosphorylation. And also, in contrast to the previous method, the simulated spectral library covering all possible singly phosphorylated forms enabled the confident site localization by SpectraST searching. The proposed approach, spectral simulation and SpectraST (SimSpectraST), was further evaluated in comparison to different localization tools.

Performance Evaluation of SimSpectraST, Mascot/phosphoRS, Sequest/Ascore, and Andromeda/PTM Score, On Synthetic Phosphopeptide Data Sets

Before comparing localization tools, scores used in SpectraST were optimized for our purpose. Since SpectraST considers the deltaDot score only between nonhomologous alternative peptides, that score was recalculated to discriminate alternative phosphorylation sites even when those sites are on the same peptide sequence. In SimSpectraST searching of the 20 synthetic phosphopeptide HCD data using SimHeLa library (see Tables 1 and 2), a recalculated deltaDot score $\geq$0.005 maximized the localization sensitivity to 67% at 1% FLR (Supporting Information Figure 9). Therefore, that score cutoff was selected as a prerequisite for the site localization, and used in combination with the total discriminant score *F*-value.

SimSpectraST searching was evaluated by comparison to other tools developed for phosphorylation site localization, including phosphoRS[22] on Mascot searching, Ascore[21] on Sequest searching, and PTM score[23] on Andromeda searching. At 1% FLR on the 20 synthetic phosphopeptide HCD data,

DOI: 10.1021/acs.jproteome.5b00050
J. Proteome Res. 2015, 14, 2348−2359

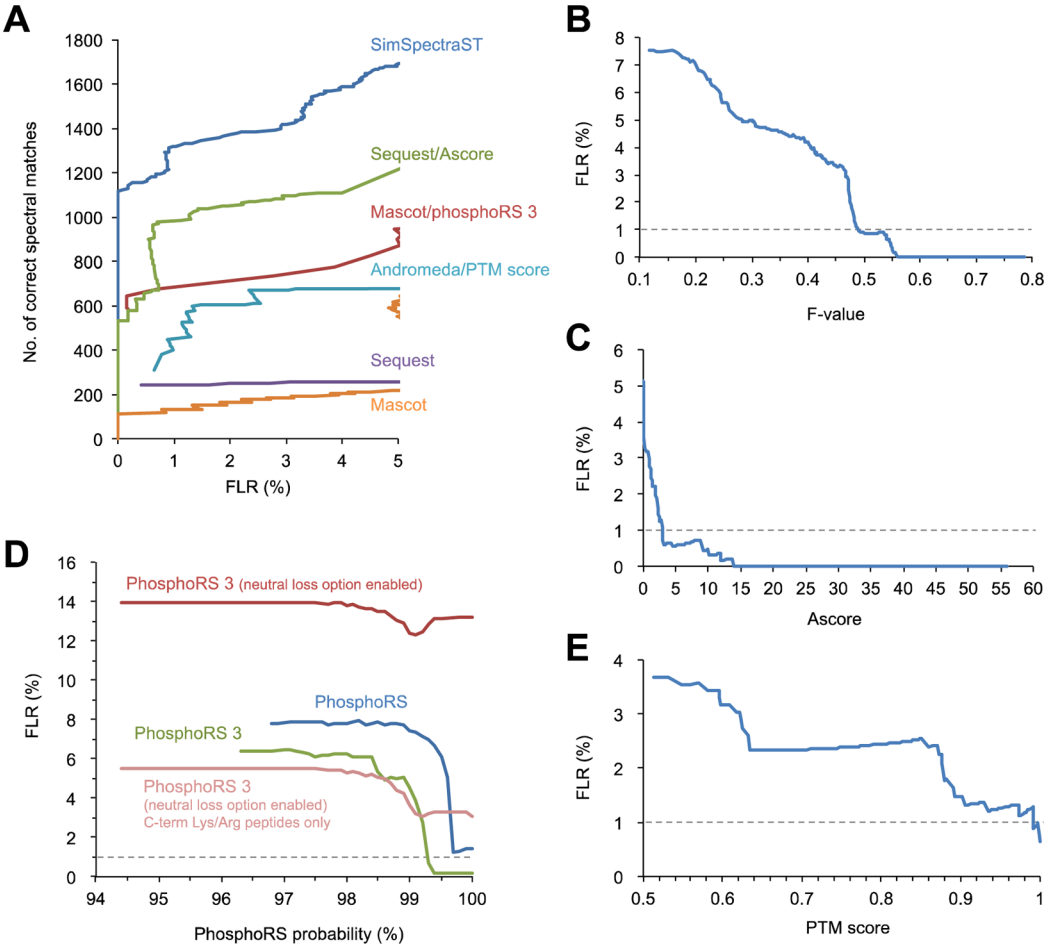


Figure 2. Phosphorylation site localization on synthetic phosphopeptides using different methods. (A) In the site localization on the 20 synthetic phosphopeptide HCD data (see Table 1), SimSpectraST was compared with 5 other tested programs/combinations, including Sequest with and without Ascore, Mascot with and without phosphoRS v3, and Andromeda with PTM score. The SimSpectraST result obtained using SimHeLa library (see Table 2) was sorted by *F*-value, after applying a recalculated deltaDot score ≥0.005 (see Supporting Information Figure 9). The neutral loss option of phosphoRS v3 was disabled (refer to panel D). SimSpectraST outperformed the other methods in the number of correct spectrum matches at 0−5% FLRs. (B−E) Score thresholds for 1% FLR were examined on these search results. (B) FLRs across SimSpectraST *F*-value are shown after applying the recalculated deltaDot cutoff. An FLR of 1% required *F*-value of 0.49. (C) FLR analysis as a function of Ascore on Sequest identifications showed Ascore of 3 as a 1% FLR cutoff. (D) FLRs as a function of phosphoRS probability on Mascot identifications are shown. PhosphoRS and its version 3 were examined. PhosphoRS v3 with the H_3PO_4 neutral loss option enabled was also tested with and without excluding peptides not containing C-terminal Lys/Arg, which generate the predominant H_2O loss (see Supporting Information Figure 5B). PhosphoRS v3 probability of 99.3% was taken as a 1% FLR cutoff. (E) FLR analysis as a function of PTM score on Andromeda identifications showed PTM score of 0.992 as a 1% FLR cutoff.

SimSpectraST searching using SimHeLa library resulted in 34% more spectral matches than other tested programs, where the second best was Sequest/Ascore, followed by Mascot/phosphoRS (v3), Andromeda/PTM score, and then Mascot and Sequest (Figure 2A and Supporting Information Table 2).

On the basis of those data, score thresholds for 1% FLR were obtained. SimSpectraST required *F*-value of 0.49 and recalculated deltaDot of 0.005 for 1% FLR (Figure 2B). Ascore of 3 was required for 1% FLR in this study (Figure 2C); however it was significantly different from the one previously reported (Ascore of 19 for 1% FLR on ion trap CID data).[21] This may be due to the difference in the fragmentation methods and/or the software versions used, since Ascore of 19 gave significantly less hits in our HCD data compared to those obtained with phosphoRS v3 cutoffs (Supporting Information Figure 10). PhosphoRS and its version 3, which is able to consider the H_3PO_4 neutral loss from phosphorylated Ser/Thr, were examined. However, considering the H_3PO_4 loss seemed to increase false localization due to the misrecognition of the H_2O neutral loss, which are frequently observed, e.g., on b-series fragment ions of peptides without the C-terminal lysine/arginine (see Figure 2D and Supporting Information Figure 5B). Therefore, the H_3PO_4 neutral loss option was not used in this study unless otherwise noted. PhosphoRS (v3, the neutral loss option disabled) probability of 99.3% was required for 1% FLR, but the other tested phosphoRS conditions did not reach 1% FLR on the used data set (Figure 2D). Also, 1% FLR required PTM score of 0.992 (Figure 2E).

The score threshold for SimSpectraST was further examined on a recently reported large HCD data set. Marx et al. synthesized >100 000 singly phosphorylated peptides and their nonphosphorylated forms by substituting phosphorylated and neighboring amino acids of 96 seed phosphopeptides, and acquired HCD and ETD spectra with an LTQ Orbitrap Velos.[25] As these data sets have been used in the evaluation of site localization methods including Mascot delta score, Mascot/

DOI: 10.1021/acs.jproteome.5b00050
J. Proteome Res. 2015, 14, 2348−2359

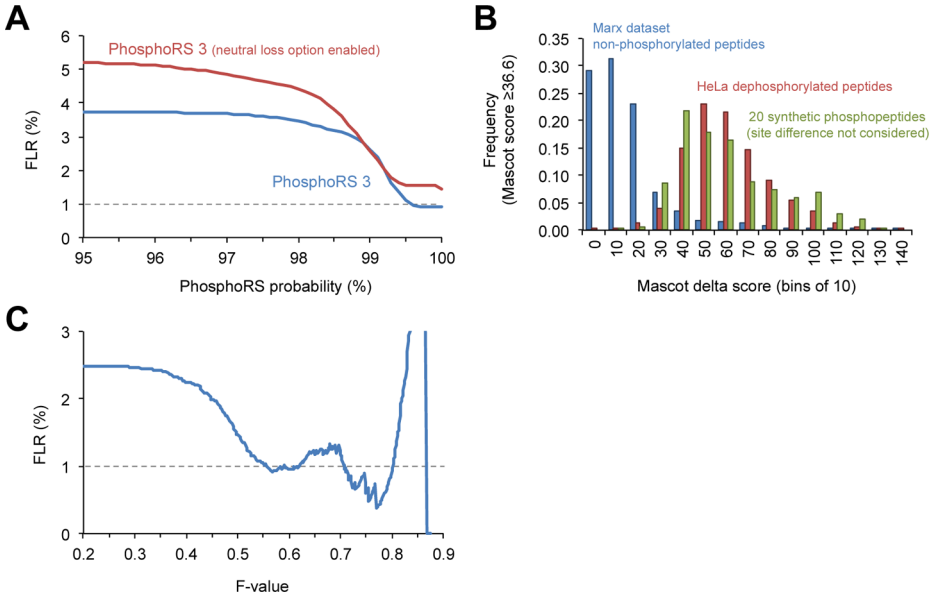


Figure 3. Evaluation of Mascot/phosphoRS and SimSpectraST on a large HCD data set of synthetic phosphopeptides. An HCD data set of >100 000 synthetic phosphopeptides and their nonphosphorylated forms generated by Marx et al.[25] (Marx data set, see Table 1) was used. (A) Mascot/PhosphoRS v3 was evaluated on the phosphopeptides of the Marx data set. On Mascot identifications at 1% FDR (target-decoy, Mascot score ≥28.2, 97 253 phosphopeptide spectral matches), phosphoRS v3 required probability of 99.6% for 1% FLR. (B) Mascot delta score at a sequence identification level was investigated. In the identifications of the nonphosphorylated counterparts from the Marx data set at 1% FDR (target-decoy, Mascot score ≥36.6), 29.1% of the spectral matches were observed with Mascot delta score of 0, and 60.4% with the delta score ≤10. In contrast, HCD data of HeLa dephosphorylated peptides and the 20 synthetic phosphopeptides (see Table 1) showed 0.2% and 0.1% of spectral matches with the delta score ≤10 in the same condition (Mascot score ≥36.6), respectively. These indicate the sequence identification ambiguity for the Marx data set. (C) SimSpectraST was evaluated on the Marx data set using SimMarx library (see Table 2). Only SimSpectraST identifications sharing sequences with the Mascot identifications at 1% FDR were taken into account (31 037 phosphopeptide spectral matches). SimSpectraST required *F*-value of 0.556 for 1% FLR, after applying recalculated deltaDot score ≥0.005.

phosphoRS (v2) and Andromeda/PTM score,[25] we also tested this large HCD data set (Marx data set, see Table 1) with Mascot/phosphoRS v3 and subsequently with SimSpectraST. On Mascot identifications at a target-decoy FDR of 1%, phosphoRS v3 required probability of 99.6% for 1% FLR, while considering the neutral loss did not reach 1% FLR also on this data set (Figure 3A). This probability threshold of 99.6% was identical to the one obtained with phosphoRS v2 on the same data set previously.[25] However, it was higher than the above-mentioned threshold obtained on the 20 synthetic phosphopeptide HCD data (99.3%, see Figure 2D), presumably due to the lower reliability of sequence identifications from the Marx data set. This data set is complex, with the exceptionally large number of homologous sequence peptides (isomeric, isobaric, and similar mass), which are difficult to be discriminated. Indeed, in the identifications of the nonphosphorylated counterparts from the Marx data set at 1% FDR, 29.1% of the spectral matches were observed with Mascot delta score of 0 (i.e., ambiguous identifications), and 60.4% with the delta score ≤10 (Figure 3B). This was significantly different from those of our HCD data sets (Table 1), as the HeLa dephosphorylated peptides and the 20 synthetic phosphopeptides (site isoforms not considered) showed 0.2% and 0.1% of the spectral matches with the delta score ≤10, respectively (Figure 3B). Nevertheless, the data set was tested with SimSpectraST. A simulated phosphopeptide spectral library was built from the nonphosphorylated peptides of the Marx data set by using stringent criteria, the delta score ≥10 and PeptideProphet probability ≥0.95 (SimMarx library, see Table 2). In SimSpectraST searching of the phosphopeptides from the Marx data set, the sequence identification was rather challenging probably due to the use of wide fragment mass tolerance by SpectraST. Therefore, only phosphopeptide spectral matches sharing the same sequences with the Mascot identifications (1% FDR) were taken into account. Consequently, SimSpectraST was able to reach 1% FLR on this large Marx data set (Figure 3C). However, as well as phosphoRS v3, SimSpectraST required higher *F*-values for 1% FLR on the Marx data set than those on the 20 synthetic phosphopeptide HCD data (Figures 2B and 3C). Particularly, since *F*-value does not discriminate homologous peptides, the result was significantly influenced by the Marx data set (Figure 3C). Despite the use of the large synthetic phosphopeptide data set, site localization on ambiguous sequence identifications is not reliable; therefore, we used the score thresholds obtained on the small, but more reliable, data set in the following analysis.

Performance Comparison of SimSpectraST with Mascot/phosphoRS and Sequest/Ascore, on HeLa Phosphopeptide Data

SimSpectraST searching was evaluated also on singly phosphorylated peptides in the HeLa phosphopeptide HCD data using SimHeLa library (see Tables 1 and 2), at an FDR of 1% estimated by PeptideProphet (Supporting Information Figure 11). The result was compared to that obtained with Mascot/phosphoRS v3 and Sequest/Ascore. Although the three identification methods (i.e., SimSpectraST, Mascot, and Sequest) agreed only on 78.4% of phosphorylation sites on shared sequence identifications, it was clearly improved to 97.3% after applying the respective 1% FLR cutoffs of the localization methods (SimSpectraST, phosphoRS v3, and Ascore) (Figure 4A, and Supporting Information Figure 12).

DOI: 10.1021/acs.jproteome.5b00050
J. Proteome Res. 2015, 14, 2348−2359

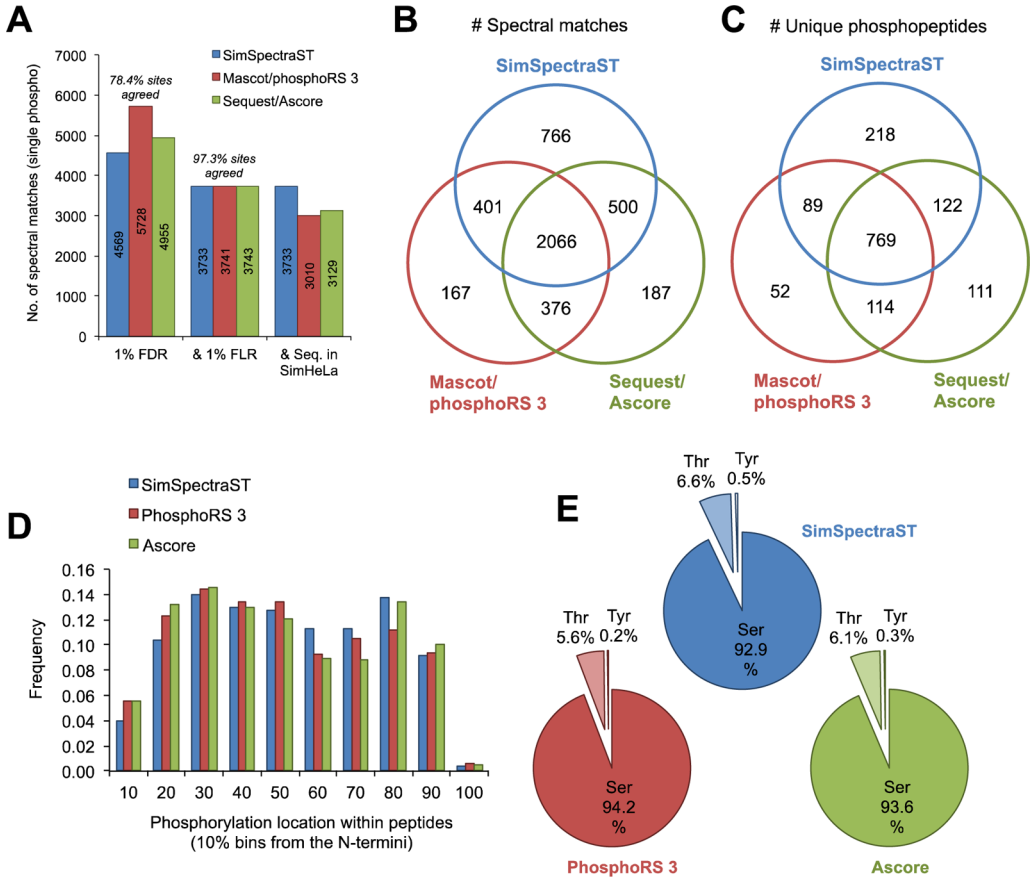


Figure 4. Phosphorylation site localization on HeLa phosphopeptides using SimSpectraST, Mascot/phosphoRS, and Sequest/Ascore. (A) From the HeLa phosphopeptide HCD data (see Table 1), singly phosphorylated peptides were identified at 1% FDR (PeptideProphet) using SimSpectraST, Mascot, and Sequest, followed by applying 1% FLR cutoffs to SimSpectraST, phosphoRS v3, and Ascore, respectively (see Figure 2 for the FLR cutoffs). SimHeLa library was used for the SimSpectraST search (see Table 2). The three identification methods agreed on 78.4% of phosphorylation sites on shared sequence identifications, which was increased to 97.3% by applying the FLR cutoffs (see Supporting Information Figure 12). SimSpectraST showed more spectral matches than the other methods when compared only for the peptide sequences covered by SimHeLa library. (B and C) In that condition, Venn diagrams show the overlap of (B) spectral matches and (C) unique phosphopeptides obtained using the three methods. (D and E) Also in that condition, no significant difference in (D) the distribution and (E) composition of phosphorylation sites was observed among the three methods.

The FLR cutoffs were applied only to peptides having >1 Ser/Thr/Tyr residues in the sequences. The SimSpectraST results at 1% FDR and 1% FLR are listed in Supporting Information Table 3, and also all the results of the three methods at 1% FDR are summarized in Supporting Information Table 4. To estimate actual FLRs in those results, all the site disagreement spectra (56 spectra) were manually interpreted. When we assume that all the agreed sites were correctly localized, estimated FLRs for SimSpectraST, phosphoRS v3, and Ascore were calculated as 1.5%, 0.63%, and 0.92%, respectively (Supporting Information Figure 12). These are well approximated to the expected FLR of 1%, supporting the validity of the cutoff scores obtained on the 20 synthetic phosphopeptide HCD data (refer to Figure 2). While the numbers of spectral matches at that condition (1% FDR, 1% FLR) were similar among the three methods (0.26% difference), SimSpectraST showed >16% more spectral matches than the others when counted only for the peptide sequences covered by SimHeLa library (Figure 4A,B). Those spectral matches by SimSpectraST resulted in 1198 unique phosphopeptides, which was >6.8% more than those by the other two methods (Figure 4C). These results support that SimSpectraST searching was sensitive in the phosphopeptide identification with simultaneous confident localization when compared in the same condition, while it has the limitation in the peptide coverage.

In that condition, the distribution (between peptide N- and C-termini) and composition (Ser, Thr, and Tyr) of the localized HeLa phosphorylation sites were similar among the three methods (Figure 4D,E). SimSpectraST localized phosphorylation sites to the N-terminal positions of tryptic peptides somewhat less frequently than phosphoRS v3 and Ascore; however, no significant difference in the site distribution was observed among the three methods (Figure 4D). In other words, the result of SimSpectraST would be as biased as those of phosphoRS v3 and Ascore. Most of tryptic peptides have the C-terminal lysine/arginine, therefore the peptide C-terminal phosphorylation sites are observed less frequently than other sites within peptide sequences. Also, tryptic cleavage tends to be inhibited by neighboring phosphorylation, which may reduce the occurrence of the peptide N-terminal phosphorylation sites. From the analytical point of view, localizing the N-terminal sites by HCD is sometimes difficult since a_1 and b_1 fragment ions are usually not observed (unless the peptides are N-terminally modified or contain N-terminal lysine/arginine) and furthermore the predominant H_2O loss on a- and b-ions may mimic the

DOI: 10.1021/acs.jproteome.5b00050
J. Proteome Res. 2015, 14, 2348−2359

H_3PO_4 loss from phosphorylated serine/threonine (see representative MS/MS spectra in Supporting Information Figure 5B). Regarding the composition of the phosphorylation sites, the three methods localized: Ser, 92.9−94.2%; Thr, 5.6−6.6%; and Tyr, 0.2−0.5% (Figure 4E).

The effect of the size of spectral libraries was investigated in SimSpectraST searching. SimHeLa library was supplemented with publically available ion trap CID spectral libraries of mouse and yeast phosphopeptides, resulting in a 7-fold larger library (SimHeLa-mouse-yeast library, see Table 2). In SimSpectraST searching of the 20 synthetic phosphopeptide HCD data (see Table 1) at 1% FLR, SimHeLa-mouse-yeast library required the same *F*-value threshold (0.49) and showed 2.1% less spectral matches than SimHeLa library (Supporting Information Figure 13A,B). In the searching of the HeLa phosphopeptide HCD data, SimHeLa-mouse-yeast library showed 95 less spectral matches to human phosphopeptides (2.5% less) than SimHeLa library. Instead, additional 205 spectral matches to mouse/yeast phosphopeptides were observed (Supporting Information Figure 13C), even though the site localization by CID-HCD spectral matching was not so reliable (see Supporting Information Figure 7B). These results indicate that, the increased size of the spectral library did not lead to significant change in the score threshold and the sensitivity.

CONCLUSION

This study shows that simulated MS/MS spectra of singly phosphorylated peptides in combination with SpectraST spectral library searching successfully leads to accurate phosphorylation localization outperforming all other available and tested programs. For preventing the false localization, a spectral library of phosphopeptides needs to contain all possible isoforms, which was readily accomplished by the simulation algorithm. The results obtained from the HeLa sample and the synthetic peptides indicate that SimSpectraST is a sensitive method for the site-specific identification of phosphopeptides. While the identification capability is currently limited due to the incomplete peptide coverage of the library, our results suggest that SimSpectraST serves as a promising orthogonal method for supplementing and/or validating the phosphorylation sites obtained by database searching and localization tools. Creating a spectral library is indeed laborious; however, that library can be used in many (10s, 100s, 1000s, ...) searches, e.g., for large-scale label-free quantifications. The simulation was developed for the HCD spectra but should also work on other beam-type MS instruments such as Q-TOFs. With some modifications, we expect the method to be applicable to ETD fragmentation. Finally, simulated phosphopeptide spectra could also be used for targeted quantification by selected reaction monitoring (SRM) in creation of transition lists and for data-independent analysis in creation of spectral libraries, in cases when particular phosphopeptides have not been observed by MS/MS before, cutting the costs of synthetic peptides. For further development, simulation of multiply phosphorylated peptides remains to be implemented.

ASSOCIATED CONTENT

Supporting Information

Supporting Information Tables 1 and 2 contain the list and data of 20 synthetic phosphopeptides, respectively. Supporting Information Tables 3 and 4 contain HeLa phosphopeptide data. Supporting Information Figures contain various detailed results. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

* E-mail: corthals@uva.nl. Tel.: +31205255406.

Author Contributions

#V.S. and S.Y.I. contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Arttu Heinonen and Pekka Haapaniemi for instrument support at the Turku Proteomics Facility, and the PRIDE team for making our data publicly accessible. This work was supported by the Academy of Finland, Nordforsk, Biocentre Finland (G.L.C), the Turku University Foundation (V.S. and S.Y.I.), PhosphoNetX, a project supported by SystemsX, and the Swiss initiative for systems biology (R.A.). V.S. acknowledges support from the Turku Centre for Computer Science graduate school. A.M. was supported by EMBO long term fellowship.

ABBREVIATIONS

AMBIC, ammonium bicarbonate; ETD, electron-transfer dissociation; FA, formic acid; FDR, false discovery rate; FLR, false localization rate; HCD, higher energy collisional dissociation; MSA, multistage activation; MS/MS, tandem mass spectrometry; SimSpectraST, spectral simulation and SpectraST; SRM, selected reaction monitoring; TCEP, tris(2-carboxyethyl)phosphine; TPP, Trans Proteomic Pipeline

REFERENCES

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198−207.

(2) Steen, H.; Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **2004**, *5* (9), 699−711.

(3) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R., III. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **2013**, *113* (4), 2343−2394.

(4) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, *7* (5), 655−667.

(5) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* **2006**, *5* (8), 1843−1849.

(6) Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **2006**, *78* (16), 5678−5684.

(7) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R. Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* **2008**, *5* (10), 873−875.

(8) Bodenmiller, B.; Malmstrom, J.; Gerrits, B.; Campbell, D.; Lam, H.; Schmidt, A.; Rinner, O.; Mueller, L. N.; Shannon, P. T.; Pedrioli, P. G.; Panse, C.; Lee, H. K.; Schlapbach, R.; Aebersold, R. PhosphoPep— phosphoproteome resource for systems biology research in Drosophila Kc167 cells. *Mol. Syst. Biol.* **2007**, *3*, 139.

(9) Bodenmiller, B.; Campbell, D.; Gerrits, B.; Lam, H.; Jovanovic, M.; Picotti, P.; Schlapbach, R.; Aebersold, R. PhosphoPep—a database of protein phosphorylation sites in model organisms. *Nat. Biotechnol.* **2008**, *26* (12), 1339−1340.

DOI: 10.1021/acs.jproteome.5b00050
J. Proteome Res. 2015, 14, 2348−2359

(10) Hummel, J.; Niemann, M.; Wienkoop, S.; Schulze, W.; Steinhauser, D.; Selbig, J.; Walther, D.; Weckwerth, W. ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites. *BMC Bioinf.* **2007**, *8*, 216.
(11) Desiere, F.; Deutsch, E.; King, N.; Nesvizhskii, A.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S.; Aebersold, R. The PeptideAtlas project. *Nucleic Acids Res.* **2006**, *34* (Database issue), D655−D658.
(12) Hu, Y.; Li, Y.; Lam, H. A semi-empirical approach for predicting unobserved peptide MS/MS spectra from spectral libraries. *Proteomics* **2011**, *11* (24), 4702−4711.
(13) Hu, Y.; Lam, H. Expanding tandem mass spectral libraries of phosphorylated peptides: advances and applications. *J. Proteome Res.* **2013**, *12* (12), 5971−5977.
(14) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76* (14), 3908−3922.
(15) Zhang, Z. Q. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal. Chem.* **2005**, *77* (19), 6364−6373.
(16) Imanishi, S. Y.; Kochin, V.; Ferraris, S. E.; de Thonel, A.; Pallari, H. M.; Corthals, G. L.; Eriksson, J. E. Reference-facilitated phosphoproteomics: fast and reliable phosphopeptide validation by microLC-ESI-Q-TOF MS/MS. *Mol. Cell. Proteomics* **2007**, *6* (8), 1380−1391.
(17) Annan, R. S.; Carr, S. A. Phosphopeptide analysis by matrix-assisted laser desorption time-of-flight mass spectrometry. *Anal. Chem.* **1996**, *68* (19), 3413−3421.
(18) Schroeder, M. J.; Shabanowitz, J.; Schwartz, J. C.; Hunt, D. F.; Coon, J. J. A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal. Chem.* **2004**, *76* (13), 3590−3598.
(19) Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **2005**, *1* (2005), 0017.
(20) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74* (20), 5383−5392.
(21) Beausoleil, S. A.; Villén, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **2006**, *24* (10), 1285−1292.
(22) Taus, T.; Köcher, T.; Pichler, P.; Paschke, C.; Schmidt, A.; Henrich, C.; Mechtler, K. Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* **2011**, *10* (12), 5354−5362.
(23) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **2006**, *127* (3), 635−648.
(24) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367−1372.
(25) Marx, H.; Lemeer, S.; Schliep, J. E.; Matheron, L.; Mohammed, S.; Cox, J.; Mann, M.; Heck, A. J.; Kuster, B. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat. Biotechnol.* **2013**, *31* (6), 557−564.
(26) Roepstorff, P.; Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* **1984**, *11* (11), 601.
(27) Biemann, K. Contributions of mass spectrometry to peptide and protein structure. *Biomed. Environ. Mass Spectrom.* **1988**, *16* (1−12), 99−111.
(28) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24* (21), 2534−2536.
(29) Lam, H.; Deutsch, E. W.; Aebersold, R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *J. Proteome Res.* **2010**, *9* (1), 605−610.
(30) Baumgardner, L. A.; Shanmugam, A. K.; Lam, H.; Eng, J. K.; Martin, D. B. Fast parallel tandem mass spectral library searching using GPU hardware acceleration. *J. Proteome Res.* **2011**, *10* (6), 2882−2888.
(31) Frese, C. K.; Altelaar, A. F.; Hennrich, M. L.; Nolting, D.; Zeller, M.; Griep-Raming, J.; Heck, A. J.; Mohammed, S. Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos. *J. Proteome Res.* **2011**, *10* (5), 2377−2388.

DOI: 10.1021/acs.jproteome.5b00050
J. Proteome Res. 2015, 14, 2348−2359