Anal. Chem. **2009,** *81,* 5794–5805

Comprehensive and Reliable Phosphorylation Site Mapping of Individual Phosphoproteins by Combination of Multiple Stage Mass Spectrometric Analysis with a Target-Decoy Database Search

Guanghui Han,[†] Mingliang Ye,*[,†] Xinning Jiang,[†] Rui Chen,[†] Jian Ren,[‡] Yu Xue,[‡] Fangjun Wang,[†] Chunxia Song,[†] Xuebiao Yao,[‡] and Hanfa Zou*[,†]

CAS Key Laboratory of Separation Sciences for Analytical Chemistry, National Chromatographic R&A Center, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China, and Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science & Technology of China, Hefei 230027, China

Since the emergence of proteomics, much attention has been paid to the development of new technologies for phosphoproteomcis analysis. Compared with large scale phosphorylation analysis at the proteome level, comprehensive and reliable phosphorylation site mapping of individual phosphoprotein is equally important. Here, we present a modified target-decoy database search strategy for confident phosphorylation site analysis of individual phosphoproteins without manual interpretation of spectra. Instead of using all protein sequences in a proteome database of an organism for the construction of a target-decoy database for phosphoproteome analysis, the composite database constructed for phosphorylation site analysis of individual phosphoproteins only included the sequences of the individual target proteins and a decoy version of a small inhomogeneous protein database. It was found that the confidence of phosphopeptide identifications could be effectively controlled when the acquired MS^2 and MS^3 spectra were searched against the above composite database followed with data processing. Because of the small size of the composite database, the computation time for the database search is very short, which allows the adoption of low-specificity proteases for protein digestion to increase the coverage of phosphorylation site mapping. The sensitivity and comprehensive phosphorylation site mapping of this approach was demonstrated by using two standard phosphoprotein samples of α-casein and β-casein, and this approach was further applied to analyze the phosphorylation of the cyclic AMP-dependent protein kinase (PKA), which resulted in the identification of 17 phosphorylation sites, including five novel sites on four PKA subunits.

Reversible protein phosphorylation is a central cellular regulatory mechanism in modulating protein activity and propagating signals within cellular pathways and networks. Conversely, abnormal phosphorylation is a cause or consequence of multiple diseases, including cancer.[1] Knowing the phosphorylated residues in proteins is fundamental for understanding the various signaling events in which they partake; therefore, much effort has been invested in trying to identify and characterize phosphorylation sites. In many cases, a protein can be phosphorylated on multiple sites, which can either act independently or synergistically when phosphorylated simultaneously. Thus, improved methods with which to comprehensively, sensitively, and reliably detect and analyze phosphorylation sites have always been sought to understand this important modification.[2−4]

Traditional methods for measuring protein phosphorylation such as mutational analysis and Edman degradation chemistry on phosphopeptides have the disadvantage of being relatively time-consuming and laborious, requiring large amounts of purified protein. Although there are a variety of methods available, mass spectrometry (MS) recently has become the primary choice for the study of protein phosphorylation because of its high sensitivity, selectivity, and speed.[5−7] Presently, most MS-based phosphoproteomics analyses adopt the "bottom-up" approach. This approach involves enzymatic cleavage of proteins, most often by trypsin, with subsequent phosphopeptide enrichment and nano-LC-MS/MS analysis to identify phosphopeptides. Even though large scale phosphoproteome analyses could presently identify ten thousands of phosphorylation sites from a single biologic sample,[8,9] mapping of phosphorylation sites for individual phosphoproteins is not comprehensive because of the extreme complexity of the pro-

* To whom correspondence should be addressed: (H. Zou) Phone: +86-411-84379610. Fax: +86-411-84379620. E-mail: hanfazou@dicp.ac.cn. (M. Ye) Phone: +86-411-84379620. Fax: +86-411-84379620. E-mail: mingliang@dicp.ac.cn.

[†] Chinese Academy of Sciences.

[‡] University of Science & Technology of China.

(1) Hunter, T. *Cell* **2000**, *100*, 113–127.
(2) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. *Cell* **2006**, *127*, 635–648.
(3) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. *Nat. Biotechnol.* **2006**, *24*, 1285–1292.
(4) Schmezle, K.; White, F. M. *Curr. Opin. Biotechnol.* **2006**, *17*, 406–414.
(5) Mann, M.; Ong, S. E.; Gronborg, M.; Steen, H.; Jensen, O. N.; Pandey, A. *Trends Biotechnol.* **2002**, *20*, 261–268.
(6) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198–207.
(7) Han, G. H.; Ye, M. L.; Zou, H. F. *Analyst* **2008**, *133*, 1128–1138.
(8) Zhai, B.; Villén, J.; Beausoleil, S. A.; Mintseris, J.; Gygi, S. P. *J. Proteome Res.* **2008**, *7*, 1675–1682.

10.1021/ac900702g CCC: $40.75
Published on Web 06/12/2009

teome sample.[10,11] For example, only two phosphorylation sites on period 2 protein could be identified by large-scale phosphoproteome analysis of the sample, while detailed analysis of the individual phosphoprotein resulted in detection of more than 20 in vivo phosphorylation sites.[12] Therefore, in order to comprehensively and reliably localize phosphorylation sites of some individual phosphoproteins, detailed analysis of a sample containing only one or a few phosphoproteins is desirable.

The most challenge step for the mapping of phosphorylation sites on individual phosphoproteins is how to confidently identify phosphopeptides. Phosphopeptide identification is based on peptide fragmentation by collisionally activated tandem mass spectrometry (MS/MS or MS^2). However, the MS^2 spectra for phosphopeptides often lack enough fragment peaks due to neural loss of H_3PO_4, and the assignment of phosphorylation sites was ambiguous in most instances when the peptides contain several potential phosphorylation sites.[13] Therefore, manual interpretation is often used to localize the phosphorylation sites.[14] However, this is a very time-consuming and labor-intensive procedure that has become impractical because data sets have grown in size. In addition, success of this strategy strongly depends on personal experience to analyze the data sets. Thus, the obtained results are typically not objective, and confidence of identification is hard to control. To circumvent these limitations, Schlosser et al.[12] have developed a novel score scheme for in-depth analysis of individual phosphoproteins. In their scoring scheme, the approach that an expert mass spectrometrist would use for manual interpretation of phosphopeptide MS^2 spectra was mimiced. It was demonstrated that their scheme was very useful in assisting phosphorylated site mapping. Because of low quality of MS^2 spectra for phosphopeptides, their scheme still lacks enough sensitivity. As supplementary to MS^2, a neutral loss peak could be further fragmented to generate MS^3 spectrum, and more fragment information could be obtained. Some phosphopeptides that could not be identified by MS^2 were successfully identified by MS^3.[15−17] MS^3 spectra were demonstrated to be beneficial for phosphoproteome analysis, especially when the peptide assignments derived from MS^2 and MS^3 were combined.[18,19] Therefore, combinational usage of MS^2 and MS^3 should also lead to more confident and more sensitive mapping of phosphorylation sites for individual phosphoproteins in a less complex sample.

Target-decoy search is a good approach for the evaluation of the confidence of peptide identification for proteome analysis.[3,20,21] After database searching against a composite protein database, including target (forward) and decoy (reversed) sequences of all proteins in the proteome of an organism, a false discovery rate (FDR) can be easily determined through the number of decoy identifications. Using the target-decoy search strategy for the acquired spectra, a data set of peptide identifications with low FDR (for example, 2%) could be easily established through postsearch filtering with easily accessible criteria. In order to circumvent labor-intensive manual validation and control the confidence of phosphopeptide identification, the target-decoy approach was successfully applied for phosphoproteome analysis. For large-scale analysis, a high-accuracy mass spectrometer incorporated with a MS^2 target-decoy search strategy[2,3] and a low-accuracy mass spectrometer (such as ion trap mass spectrometer) with a MS^2/MS^3 target-decoy search strategy[18,19,22] have been reported to obtain high confident phosphopeptide identification and precise site location without manual validation. However, to the best of our knowledge, a MS^2/MS^3 target-decoy search strategy for comprehensive mapping of phosphorylation sites on individual phosphoproteins has not been reported.

Here, we present a methodology for confident phosphorylation site analysis of individual phosphoproteins by a MS^2/MS^3 target-decoy strategy. Instead of using all protein sequences in a proteome database of an organism for the construction of a target-decoy database for phosphoproteome analysis, the composite database constructed for phosphorylation site analysis of individual phosphoproteins only included the sequences of the target individual protein(s) and a decoy version of a small inhomogeneous protein database. The effectiveness of using the above small composite database to control the confidence of phosphopeptide identifications for the analysis of individual phosphoproteins was demonstrated by analysis of phosphorylation sites of α-casein and β-casein. Because of the extremely slow database searching when low-specificity proteases are applied, phosphoproteome analysis is limited to using high-specific proteases like trypsin for digestion of proteins. However, the composite database for phosphorylation site mapping of individual proteins is much smaller, and the database search is much faster. Thus, low-specificity proteases could be applied to increase the coverage of phosphorylation site mapping. In combination with a multiprotease digestion approach, phosphorylation sites of α-casein and β-casein can be comprehensively, sensitively, and reliably detected and located. It was further applied to analyze phosphorylation of the cyclic AMP-dependent protein kinase (PKA), and 17 phosphorylation sites were confidently located on four PKA subunits. As the confidence of phosphopeptide identification could be easily controlled with the target-decoy approach, no manual inter-

(9) 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. *Mol. Syst. Biol.* **2007**, *3*, 11.
(10) Graham, M. E.; Anggono, V.; Bache, N.; Larsen, M. R.; Craft, G. E.; Robinson, P. J. *J. Biol. Chem.* **2007**, *282*, 14695–14707.
(11) Craft, G. E.; Graham, M. E.; Bache, N.; Larsen, M. R.; Robinson, P. J. *Mol. Cell. Proteomics* **2008**, *7*, 1146–1161.
(12) Schlosser, A.; Vanselow, J. T.; Kramer, A. *Anal. Chem.* **2007**, *79*, 7439–7449.
(13) Edelson-Averbukh, M.; Pipkorn, R.; Lehmann, W. D. *Anal. Chem.* **2007**, *79*, 3476–3486.
(14) Schlosser, A.; Vanselow, J. T.; Kramer, A. *Anal. Chem.* **2005**, *77*, 5243–5250.
(15) Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villen, J.; Li, J. X.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 12130–2135.
(16) Olsen, J. V.; Mann, M.H *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 13417–13422.
(17) Lee, J.; Xu, Y.; Chen, Y.; Sprung, R.; Kim, S. C.; Xie, S.; Zhao, Y. *Mol. Cell. Proteomics* **2007**, *6*, 669–676.
(18) Jiang, X.; Han, G.; Feng, S.; Jiang, X.; Ye, M.; Yao, X.; Zou, H. *J. Proteome Res.* **2008**, *7*, 1640–1649.
(19) Ulintz, P. J.; Bodenmiller, B.; Andrews, P. C.; Aebersold, R.; Nesvizhskii, A. I. *Mol. Cell. Proteomics* **2008**, *7*, 71–87.
(20) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207–214.
(21) Lu, B. W.; Ruse, C.; Xu, T.; Park, S. K.; Yates, J. *Anal. Chem.* **2007**, *79*, 1301–1310.
(22) Han, G. H.; Ye, M. L.; Zhou, H. J.; Jiang, X. N.; Feng, S.; Jiang, X. G.; Tian, R. J.; Wan, D. F.; Zou, H. F.; Gu, J. R. *Proteomics* **2008**, *8*, 1346–1361.

pretation of MS spectra is required, which allows this approach to be used more easily and simply.

EXPERIMENTAL SECTION

Chemicals and Materials. All water used in this experiment was prepared using a Milli-Q system (Millipore, Bedford, MA). A $ZipTip_{C18}$ pipet tip was purchased from Millipore. Dithiothreitol (DTT), ammonium bicarbonate (NH_4HCO_3), and iodoacetamide (IAA) were all purchased from Bio-Rad (Hercules, CA). Formic acid (FA) and acetonitrile (ACN) were obtained from Aldrich (Milwaukee, WI). Urea, trifluoroacetic acid (TFA), sodium chloride (NaCl), α-casein, β-casein, thermolysin, trypsin (TPCK-treated, proteomics grade), and cyclic AMP-dependent protein kinase (from bovine heart) were all purchased from Sigma (St. Louis, MO); elastase, proteinase K (PCR grade), and endoproteinase Glu-C (sequencing grade) were from Roche (Mannheim, Germany). All chemicals were of analytical grade except acetonitrile, which was of HPLC grade.

Proteolytic Cleavage. For α-casein and β-casein, a total of 25 μg of protein was diluted to 100 μL with 0.1 M NH_4HCO_3 (pH 8), and then divided into 5 aliquots. About 0.2 μg of each protease was used for digestion, respectively. The digestions with trypsin, elastase, proteinase K, Glu-C, and thermolysin were performed overnight at 37 °C in 0.1 M NH_4HCO_3 (pH 8) for 18 h. All digests were dried in a vacuum concentrator and redissolved in 20 μL of 80% ACN, 6% TFA and then subjected to phosphopeptide enrichment.

For digestion of a cyclic AMP-dependent protein kinase (PKA) sample, a total of 100 μg of protein was diluted to 20 μL with a solution containing 8 M urea and 50 mM Tris-HCl at pH 8.3 and then divided into 5 aliquots. After that, 0.4 μL of 1 M DTT was added to each solution. The protein solutions were incubated at 56 °C for 45 min, and then 2 μL of 1 M IAA was added and incubated for an additional 30 min at room temperature in darkness. The protein solutions were diluted by 10-fold with 0.1 M NH_4HCO_3 (pH 8) for trypsin, elastase, proteinase K, Glu-C, and thermolysin digestion. About 0.8 μg of each protease was used for digestion. The digestions with trypsin, elastase, proteinase K, Glu-C, and thermolysin were performed overnight at 37 °C in 0.1 M NH_4HCO_3 (pH 8) for 18 h. After incubation, 2.5 μL of each digest was dispensed into a clean tube, and then desalted with $ZipTip_{C18}$ as product's instruction for protein identification by LC-MS^2, respectively. Another 30 μL of each digest was dried in a vacuum concentrator and redissolved in 40 μL of 80% ACN, 6% TFA and then subjected to phosphopeptides enrichment.

Enrichment of Phosphopeptides. Immobilized titanium ion affinity chromatography (Ti^{4+}-IMAC) using phosphonate groups as chelating groups is a new generation of IMAC with high specificity for phosphopeptides.[23] Phosphopeptides in the above peptide mixtures were separately enriched by Ti^{4+}-IMAC as follows. The peptide mixture was first incubated with 10 μL of Ti^{4+}-IMAC beads (homemade, 10 mg mL^{-1}) in a loading buffer (80% ACN, 6% TFA) with a vibration of 30 min. The supernatant was removed after centrifugation, and the beads with captured phosphopeptides were washed with 50 μL of two washing buffers (50% ACN, 6% TFA containing 200 mM NaCl as washing buffer 1; 30% ACN, 0.1% TFA as washing buffer 2). The bound phosphopeptides were then eluted with 20 μL of 10% $NH_3 \cdot H_2O$ under sonication for 10 min. After centrifugation at 20000 g for 5 min, the supernatant was collected and lyophilized to dryness for phosphorylation analysis by LC-MS^2-MS^3.

Mass Spectrometric Analysis. Nano-LC-MS^2-MS^3 was performed on a nano-RPLC-MS/MS system. A Finnigan surveyor MS pump (Thermo Electron Finnigan, San Jose, CA) was used to deliver the mobile phase. For the capillary separation column, one end of the fused silica capillary (75 μm i.d. × 120 mm length) was manually pulled to a fine point, ~5 μm, with a flame torch. The column was in-house packed with C_{18} AQ beads (5 μm, 120 Å) from Michrom BioResources (Auburn, CA) using a pneumatic pump. The nano-RPLC column was directly coupled to a LTQ linear ion trap mass spectrometer from Thermo Finnigan with a nanospray source. The mobile phase consisted of mobile phase A, 0.1% formic acid (v/v) in H_2O, and mobile phase B, 0.1% (v/v) formic acid in acetonitrile.

The samples were manually loaded onto the C_{18} capillary column using a 75 μm i.d. × 220 mm length empty capillary as sample loop first, and then the reversed phase gradient was executed from 5% to 35% mobile phase B in 60 min at about 200 nL/min. A Finnigan LTQ linear ion trap mass spectrometer equipped with an ESI nanospray source was used for the MS experiment with an ion transfer capillary at 180 °C, and a voltage of 1.8 kV was applied to the cross. The LTQ instrument was operated in positive ion mode. Normalized collision energy was 35%. System control and data collection were done by Xcalibur software version 1.4. For protein identifications of PKA samples, one microscan was set for each MS and MS^2 scan. All MS and MS^2 spectra were acquired in the data-dependent mode. The mass spectrometer was set such that one full MS scan was followed by six MS^2 scans on the six most intense ions. The Dynamic Exclusion was set as follows: repeat count 2, repeat duration 30 s, and exclusion duration 90 s. For phosphorylation analysis of all samples, the mass spectrometer was set so that one full MS scan was followed by three MS^2 scans and three neutral loss MS^3 scans with the following Dynamic Exclusion settings: repeat count 2, repeat duration 30 s, exclusion duration 60 s. The detection of phosphopeptides was performed in which the mass spectrometer was set as a full scan MS followed by three data-dependent MS^2. A subsequent MS^3 spectrum was automatically triggered when one of the 10 most intense peaks from the MS^2 spectrum corresponded to a neutral loss event of 98, 49, and 32.7 ± 1 Da for the precursor ion with 1+, 2+, 3+ charge states, respectively.

Database Searching and Data Analysis. The peak lists for MS^2 and MS^3 spectra were generated from the raw data by Bioworks 3.2 (Thermo Electron) with the following parameters: mass range, 600−3500 Da; intensity threshold, 1000; precursor ion tolerance, 1.4 Da; group scan, 1; minimum group count, 1; and minimum ion count, 10.

For identification of proteins from PKA samples, the acquired MS^2 spectra were searched using Sequest (version 0.27) against a composite database including a bovine protein database and its reversed version with the following parameters: precursor−ion mass tolerance, 2 Da; fragment−ion mass tolerance, 1 Da;

(23) Yu, Z. Y.; Han, G. H.; Ye, M. L.; Sun, S. T.; Jiang, X. N.; Chen, R.; Wang, F. J.; Wu, R. A.; Zou, H. F. *Anal. Chim. Acta* **2009**, *636*, 34–41.

Table 1. Cleavage Sites of the Proteases

enzyme name	offset	cleavage sites	sites without cleavage
Glu-C	after	E	P
trypsin	after	KR	P
elastase	after	ALIVGS	−
thermolysin	before	LFIVMA	P
proteinase K		−	−

enzyme, set as shown in Table 1; missed cleavages, 2; and static modification, Cys (+57). Dynamic modifications were set for oxidized Met (+16). The bovine database was a bovine proteome sequence database (ipi.BOVIN.v3.32.fasta) from the European Bioinformatics Institute, which included 32947 entries (ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/). For identification of proteins, the following criteria were used: cross-correlation values (Xcorr) ≥ 2.0, 2.5, and 3.8 for singly, doubly, and triply charged peptides,[24] respectively, and increases in the values of ΔCn until FDR ≤ 2%.

For phosphorylation analysis, the MS^2 and MS^3 spectra were searched using Sequest (version 0.27) against a composite database, including α-S1-casein, α-S2-casein, *β*-casein sequences (or sequences of identified background proteins or PKA subunits for PKA samples), and a reversed yeast database (1000 entries as the decoy database) with the following parameters: precursor−ion mass tolerance, 2 Da; fragment−ion mass tolerance, 1 Da; enzyme, set as shown in Table 1; missed cleavages, 2; and static modification, none for casein and Cys (+57) for PKA. For searching MS^2 data, dynamic modifications were set for oxidized Met (+16), phosphorylated Ser, Thr, and Tyr (+80). For searching MS^3 data, besides the above set, dynamic modifications were also set for water loss on Ser and Thr (−18). For phosphopeptides identified by MS^2, the following criteria were used: Xcorr ≥2.0, 2.5, and 3.8 for singly, doubly, and triply charged peptides,[24] respectively, and increases in the values of ΔCn until FDR ≤ 2% or minimum FDR. For phosphopeptide identification by matching the assigned sequences derived from MS^2 and MS^3 data, a homemade software named APIVASE[18] (automatic phosphopeptide identification validating algorithm for Sequest) was applied to validate the identifications. APIVASE is available free for academic users from http://bioanalysis.dicp.ac.cn/proteomics/software/APIVASE.html. This approach was termed the MS^2/MS^3 target-decoy database search approach or MS^2/MS^3 approach in short. Briefly, there are five steps in the MS^2/MS^3 approach: (1) evaluation of the charge state to remove invalid MS^2/MS^3 pairs, (2) performing MS^2 and MS^3 target-decoy database searches separately, (3) reassignment of the peptide scores in Sequest output to generate a list of peptide identifications for pair of MS^2/MS^3 spectra, (4) filtering candidate phosphopeptides with new defined parameters (Rank'm, ΔCn'm and Xcorr's) to achieve phosphopeptide identification with specific FDR, and (5) the phosphorylation site localizations were determined by Tscore as described by Jiang et al.[18] In this study, to achieve FDR ≤ 2%, cutoff filters such as Rank'm, ΔCn'm, and Xcorr's were used to filter the data.

Table 2. Phosphorylation Sites of α-Casein Identified by Different Approaches

		MS^2/MS^3					MS^2
α-casein		trypsin	Glu-C	elastase	thermolysin	proteinase K	trypsin
S1	S56[a,b]	√	√		√	√	√
	S61[a,b]	√	√		√	√	√
	S63[a,b]	√	√		√	√	√
	T64[b]	√	√		√	√	√
	S79[a]			√			
	S81[a]						
	S82[a]						
	S83[a]						
	S90[a]			√			
	S103[c]	√			√		√
	S130[a]	√		√		√	√
S2	S23[a]	√					
	S24[a,b]	√					
	S25[a,b]	√					
	S28[b]	√					
	S31[a,b]				√		
	S46[a]	√	√				√
	S71[a]	√					
	S72[a]	√					
	S73[a]	√					
	S76[a]	√					
	S144[a,b]	√			√	√	
	T145[b]					√	
	S146[a,b]	√				√	
	S150[b]						
	T153[c]		√				
	S158[a]	√	√			√	

[a] Phosphorylation site information from ExPasy (http://www.expasy.org). [b] Phosphorylation site information from Phospho.ELM (http://phospho.elm.eu.org). [c] Phosphorylation sites localized in this study but not reported previously.

Table 3. Phosphorylation Sites of *β*-Casein Identified by Different Approaches

	MS^2/MS^3					MS^2
β-casein	trypsin	Glu-C	elastase	proteinase K	thermolysin	trypsin
S30[a,b]				√	√	
S32[a,b]				√	√	
S33[a,b]				√	√	
S34[a,b]				√	√	
S37[b]						
S50[a]	√	√	√	√	√	√
T56[b]	√		√		√	√
S111[c]		√				
S137[c]		√				
S139[b]		√				
S181[c]					√	

[a] Phosphorylation site information from ExPasy (http://www.expasy.org). [b] Phosphorylation site information from Phospho.ELM (http://phospho.elm.eu.org). [c] Phosphorylation sites localized in this study but not reported previuosly.

RESULTS AND DISCUSSION

Because of the well-characterized phosphorylation sites, two standard phosphoprotein samples, α-casein (P02662 and P02663) and *β*-casein (P02666), were chosen to test our methodology. In order to evaluate the performance of the phosphorylation site analysis, four standard measurements of accuracy (*Ac*), sensitivity (*Sn*), specificity (*Sp*), and the Mathew correlation coefficient (*MCC*) were used.[25] In this work, the known phosphorylation sites of casein from ExPasy (http://www.expasy.org) and Phospho.ELM[26] (http://phospho.e-

lm.eu.org) were regarded as positive sites (see Table 2 for the phosphorylation sites of α-casein and Table 3 for the phosphorylation sites of β-casein), while all the other (S, T, and Y) sites in the sequences of casein were regarded as negative sites. For the sites which were identified as positive, known phosphorylation ones were defined as true positives (TP), while the others were defined as false positives (FP). For the sites that were identified as negative, real positive sites were defined as false negatives (FN), while the others were called true negatives (TN). Four standard measurements of *Ac*, *Sn*, *Sp*, and *MCC* were defined as follows[25]

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$Ac = \frac{TP + TN}{TP + FP + TN + FN}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

Sn and *Sp* illustrate correct identification ratios of positive and negative sites, respectively, and *Ac* illustrates correct identification ratios of positive and negative sites. Larger values of *Sn*, *Sp*, and *Ac* stand for more correct identification, in other words, better performance for phosphorylation site localization. However, when the number of positive and negative data differ too much from each other, *MCC* should be calculated to assess the identification performance. The value of *MCC* ranges from −1 to 1, and larger *MCC* values also stands for better identification performance.[25]

The MS^2 spectra for phosphopeptides often lack enough fragment peaks due to neural loss of H_3PO_4, and manual interpretation is used to verify phosphopeptide identification for mapping of phosphorylation sites in individual phosphoproteins.[27] Only expert mass spectrometrists could effectively identify the phosphopeptides via manual interpretation. Even worse, confidence of the identifications is unknown, and results are not objective. The target-decoy database search is a popular approach for controlling the confidence of peptide identification in proteome analysis.[3,20,21] In this study, the target-decoy database search approach was applied to control the confidence of phosphopeptide identifications for individual phosphoproteins. In proteome analysis, the composite database for database search was constructed by inclusion of target (forward) and decoy (reversed) sequences of proteins presented in the proteome of an organism. However, for phosphorylation analysis of individual proteins in this study, a composite database was constructed by inclusion of sequences of proteins presented in the sample (target proteins) and a decoy version of a large enough inhomogeneous database. In the case of analysis of phosphorylation sites on α-casein and β-casein, target proteins were α-casein and β-casein, and decoy proteins were 1000 reversed sequences of yeast proteins. As the decoy database was much larger than the database of target proteins, any peptide hits from decoy database were likely to be random hits. Thus, all peptide assignments corresponding to target proteins could be considered as correct identifications and that sequences from the decoy database were incorrect. Therefore, the confidence of peptide identification could be expressed by FDR, which was calculated by the following equation: FDR = decoy/(target + decoy). In this study, the confidence of peptide identifications was controlled by adjusting suitable database search scores to achieve FDR ≤ 2% or minimum FDR, if FDR ≤ 2% was not achievable. Phosphorylation site localizations on phosphopeptides were further determined by Tscore, which was described by Jiang et al.[18]

Table 4. Comparison Accuracy (*Ac*), Sensitivity (*Sn*), Specificity (*Sp*), and Mathew Correlation Coefficient (*MCC*) of Phosphorylation Site Identifications for α-Casein and β-Casein by Different Approaches[a]

	multiproteases	trypsin	
α-casein	MS^2/MS^3	MS^2/MS^3	MS^2
FDR	1.26%	1.90%	8.72%
TP	21	17	5
FP	2	1	1
FN	4	8	20
TN	50	51	51
Sn	84.00%	68.00%	20.00%
Sp	96.15%	98.08%	98.08%
Ac	92.21%	88.31%	72.73%
MCC	82.00%	73.11%	31.58%

	multiproteases	trypsin	
β-casein	MS^2/MS^3	MS^2/MS^3	MS^2
FDR	<1.00%	<1.00%	10.68%
TP	7	2	2
FP	3	0	0
FN	1	6	6
TN	18	21	21
Sn	87.50%	25.00%	25.00%
Sp	85.71%	100.00%	100.00%
Ac	86.21%	79.31%	79.31%
MCC	68.85%	44.10%	44.10%

[a] TP, true positives; FP, false positives; FN, false negatives; TN, true negatives; *Ac*, accuracy; *Sn*, sensitivity; *Sp*, specificity; and *MCC*, Mathew correlation coefficient.

Table 5. GPS 2.0 Screening of Phosphorylation Sites of α-casein and β-casein Localized in this Study

		GPS 2.0 prediction	
proteins	sites	threshold	kinase
α-S1-casein	S103[a]	high	MAPK11
α-S2-casein	T153[a]	high	ILK
β-casein	S111[a]	high	PLK
	S137[a]	high	ATM
	S181[a]	high	ATM

[a] Phosphorylation sites localized in this study but not reported previously.

MS^2 Target-Decoy Approach. The majority of phosphorylation site mapping studies were based on MS^2. The MS^2 target-

(24) Jiang, X. N.; Jiang, X. G.; Han, G. H.; Ye, M. L.; Zou, H. F. *BMC Bioinf.* **2007**, *8*, 323.
(25) Xue, Y.; Ren, J.; Gao, X. J.; Jin, C. J.; Wen, L. P.; Yao, X. B. *Mol. Cell. Proteomics* **2008**, *7*, 1598–1608.
(26) Diella, F.; Gould, C. M.; Chica, C.; Via, A.; Gibson, T. J. *Nucleic Acids Res.* **2008**, *36*, D240−D244.
(27) Feng, S.; Ye, M. L.; Zhou, H. J.; Jiang, X. G.; Jiang, X. N.; Zou, H. F.; Gong, B. L. *Mol. Cell. Proteomics* **2007**, *6*, 1656–1665.

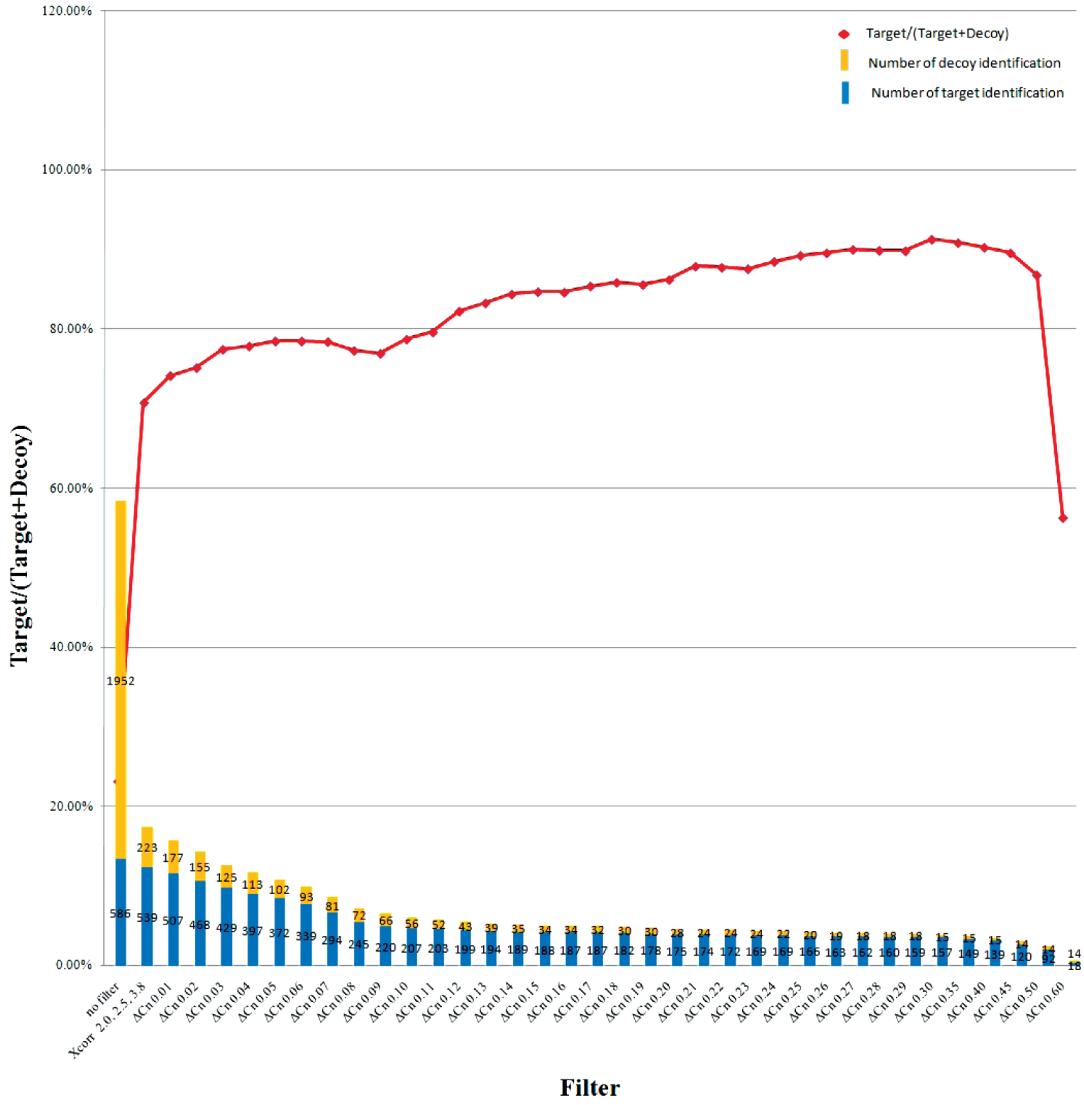


Figure 1. Number of target and decoy identifications and the percentage of target identifications using the MS^2 target-decoy strategy with Xcorr ≥2.0, 2.5, and 3.8 for singly, doubly, and triply charged peptides with increases in the value of ΔCn for all charged peptides step by step. Sample: Phosphopeptides enriched from tryptic digests of α-casein.

decoy approach refers to using only MS^2 spectra for a database search against the composite database in this study. To evaluate this approach, phosphopeptides in tryptic digests of two individual phosphoproteins, i.e. α-casein and β-casein, were enriched separately by Ti^{4+}-IMAC followed by LC-MS^2-MS^3 analysis. The acquired MS^2 spectra were then searched against the composite database. As shown in Figure 1, a total of 2538 peptide identifications, including 586 target identifications (peptide identifications from α-S1-casein or α-S2-casein), and 1952 decoy identifications (identifications of reversed yeast sequences) were obtained for α-casein without setting of cutoff filter. To improve the confidence of peptide identifications, the cutoff scores should be set to discriminate the correct hits from random hits. Different combinations of two cutoff scores, i.e., Xcorr and ΔCn, were used to filter phosphopeptide identifications: Xcorr ≥2.0, 2.5, and 3.8 for singly, doubly, and triply charged peptides, respectively, and increasing values of ΔCn for all peptides. It is shown in Figure 1 that FDR decreases with increases in ΔCn cutoff values from 0.00 to 0.30. This is reasonable as the stricter the filter criteria the more confident the peptide identification. However, with further increases in ΔCn cutoff values, FDR increased, and a significant fraction of decoy identifications remained even when the ΔCn cutoff value was as high as 0.60. The above results indicated that the MS^2 target-decoy approach cannot effectively remove random hits. The minimum FDR of 8.72% was achieved when the ΔCn cutoff value was 0.30, and the filter criteria is very strict for Xcorr and ΔCn in unmodified peptide identification of proteome analysis with the Sequest algorithm. At this FDR value, the numbers of target and decoy identifications for α-casein samples were 157 and 15, respectively. All of the 157 target peptide identifications (i.e., peptides derived from α-casein) were phosphorylated peptides. Because FDR ≤ 2% cannot be achieved here, phosphorylation sites of α-casein were determined by these phosphopeptides though their identifications are not confident enough. Among 25 known phosphorylation sites on α-casein, only 5 sites (true positive) were localized with one false positive hit. The sensitivity (*Sn*) for the phosphorylation site analysis was only 20.00%. Other measurements, *Sp* (98.08%), *Ac* (72.73%), and *MCC* (31.58%),

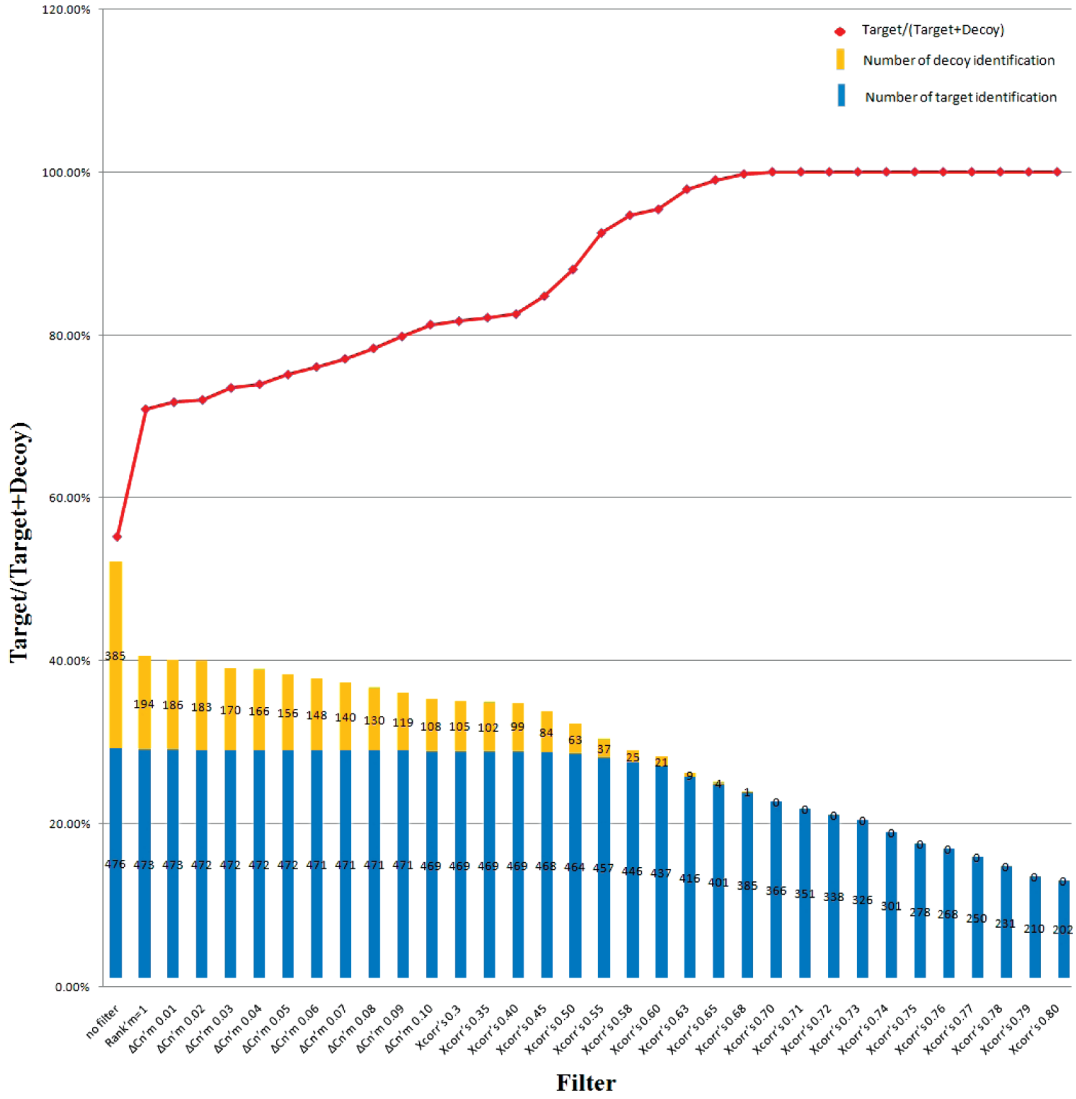


Figure 2. Target and decoy identifications and the percentage of target identifications using the MS^2/MS^3 target-decoy strategy with increases in the value of Rank'm = 1, ΔCn'm, and Xcorr's for all charged peptides step by step. Sample: Phosphopeptides enriched from tryptic digests of α-casein.

are also given in Table 4. The same procedure was applied to map phosphorylation sites of β-casein. It was found that FDR ≤ 2% could not be achieved either, and the minimum FDR was 10.68%. At this FDR value, only two sites were localized among eight known phosphorylation sites. The four measurements are also given in Table 4. Low sensitivity (*Sn* of 25.00%) was also observed for mapping of the phosphorylation sites on β-casein. From the above results, it can be concluded that the MS^2 target-decoy search strategy cannot provide confident peptide identification and also lacks enough sensitivity for phosphorylation site mapping. This is because of the poor quality of the MS^2 spectra, which suppresses phosphopeptide matching scores assigned by the current database searching algorithm.[3,28]

MS^2/MS^3 Target-Decoy Approach. We have developed an approach termed the MS^2/MS^3 target-decoy strategy for phosphoproteome analysis of HeLa cells.[18] In the strategy, MS^2 and MS^3 spectra acquired by LC-MS^2-MS^3 analysis were first verified to be valid MS^2/MS^3 pairs for phosphopeptides on the basis of their charge state and neutral loss peak. Then MS^2 and MS^3 spectra in the valid pairs were searched separately against a composite database, including forward and reversed sequences of a protein database. Only the phosphopeptides identified by both MS^2 and its corresponding MS^3 were accepted for further filtering, which greatly improved the reliability in phosphopeptide identification. It was found that sensitivity was significantly improved in the MS^2/MS^3 strategy as the number of identified phosphopeptides was 2.5 times of that obtained by a conventional filter-based MS^2 approach. Because of the use of the target-decoy database, FDR of the identified phosphopeptides could be easily determined, and no manual validation was required. In this work, the MS^2/MS^3 strategy was applied to analyze phosphorylation sites for individual phosphoproteins instead of a proteome sample, and a much smaller composite database containing only a few target proteins and 1000 yeast proteins with reversed sequences were used.

(28) DeGnore, J. P.; Qin, J. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 1175–1188.

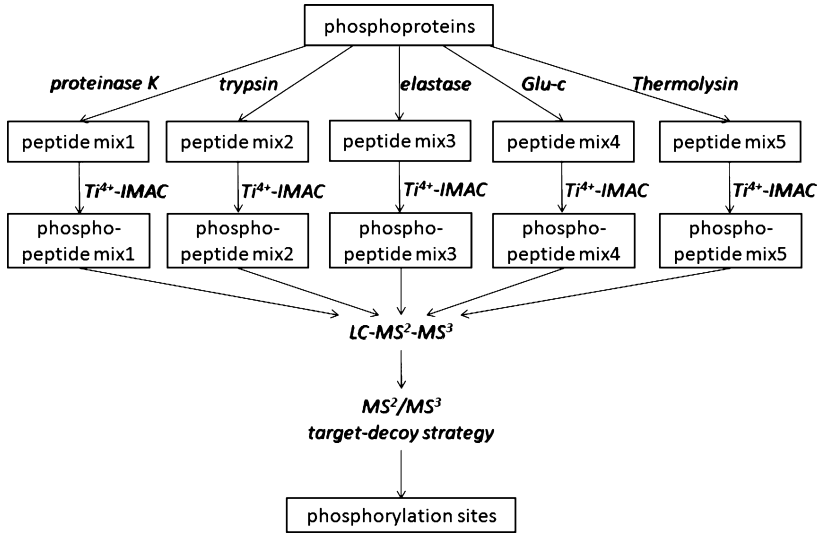
Tryptic digests of the same phosphoprotein samples, i.e., α-casein and β-casein, were used to evaluate the performance of the MS^2/MS^3 approach. The MS^2 and MS^3 spectra acquired by LC-MS^2-MS^3 analysis of enriched phosphopeptides were processed with the MS^2/MS^3 approach by a homemade software named APIVASE.[18] In the MS^2/MS^3 approach, a few new defined scores, i.e., Rank'm, ΔCn'm, and Xcorr's, were used to adjust the confidence of peptide identifications. These new scores were derived from the corresponding scores for a phosphopeptide identified by MS^2 and MS^3. As shown in Figure 2, a total of 861 peptide identifications, including 476 target identifications (peptide identifications from α-S1-casein or α-S2-casein) and 385 decoy identifications (identifications of reversed yeast protein sequences) were obtained after processing with the MS^2/MS^3 approach when no filter was used. Then different combinations of the three cutoff scores were used to filter phosphopeptide identifications using Rank'm = 1, further increasing the value of ΔCn'm until ΔCn'm ≥ 0.1, and then further increasing the value of Xcorr's. It was found that when cutoff scores were increased step by step, decoy identifications were sharply decreased and finally disappeared totally. This indicated that random hits could be effectively removed in the MS^2/MS^3 approach. When the values of the cutoff filters for the MS^2/MS^3 target-decoy analysis were set as Rank'm = 1, ΔCn'm ≥ 0.1, and Xcorr's ≥ 0.631, the result was that the number of target identifications was 414, the number of decoy identifications was 8, and the FDR was 1.90%. At this FDR value, the four measurements for phosphorylation site mapping of α-casein were *Sn* (68.00%), *Sp* (98.08%), *Ac* (88.31%), and *MCC* (73.11%) (Table 4). Compared with the MS^2 target-decoy analysis, sensitivity (*Sn*) was increased sharply from 20.00% to 68.00%. The MS^2/MS^3 approach resulted in localization of 17 sites among 25 known sites, while the MS^2 approach only resulted in localization of five sites. Besides, *Sn*, *Ac*, and *MCC* were all improved, which indicated a better performance of phosphorylation site mapping with the MS^2/MS^3 approach. Confident identification of phosphopeptides from β-casein could also be achieved by the MS^2/MS^3 approach, and FDR could be adjusted to ≤2%. The four measurements for phosphorylation site mapping of β-casein are given in Table 4. Though the confidence of phosphopeptide identification was improved significantly, the performance of phosphorylation site mapping was not improved. This is mostly because four of the eight known sites are presented in one tryptic peptide (ELEELNVPGEIVEpSLpSpSpSEESITR, MW 2965.16). Analysis of this quadruply phosphorylated peptide is extremely difficult by LC-MS/MS,[29,30] and it was not identified in this study either. The localized phosphorylation sites by the MS^2 approach and MS^2/MS^3 approach for α-casein and β-casein are given in Table 2 and Table 3. Though the quadruply phosphorylated tryptic peptide in β-casein was not identified, two quadruply phosphorylated tryptic peptides (NANEEEYSIGpSpSpSEEpSAEVATEEVK and NTMEHVpSpSpSEEpSIISQETYKQEK), one doubly phosphorylated tryptic peptide (EQLpSTpSEENSK), and three singly phosphorylated tryptic peptides (TVDMEpSTEVFTK, TVDMEpSTEVFTKK, and KTVDMEpSTEVFTKK) in α-casein were successfully identified by the MS^2/MS^3 approach, which led to localizing an additional 11 phosphorylation sites compared to the MS^2 approach. It is shown in Table 4 that few false positive localized phosphorylation sites were observed with this approach, which indicated that the controlling of confidence by target-decoy database searching with the small composite database is effective.

The improved confidence for phosphopeptide identifications by the MS^2/MS^3 approach is mainly attributed to two reasons. The first reason is that only valid MS^2/MS^3 pairs are submitted to the database search. MS^2 without MS^3 or with invalid MS^3 are removed before the database search, which eliminates random hits caused by these spectra. The second reason is that only phosphopeptide identified by both MS^2 and MS^3 are considered as valid identification, which also significantly reduces random hits. Because of the improved confidence, the MS^2/MS^3 approach allowed more poor spectra for identification of phosphopeptides if both MS^2 and MS^3 spectra were available for the phosphopeptide. For example, a singly phosphorylated peptide (KTVDMEpSTEVFTKK, triply charged) from α-casein could not be identified by the MS^2 approach because its Xcorr (2.79) and ΔCn (0.10) for the MS^2 database search did not pass the filter criteria. However, it can be effectively identified by the MS^2/MS^3 approach as the neutral loss MS^3 of the MS^2 can also match this triply charged peptide with Xcorr (3.64) and ΔCn (0.14). Though the database scores assigned for this phosphopeptide were not high for both MS^2 and MS^3, the combination of these information (Rank'm = 1, ΔCn'm = 0.396, and Xcorr's = 0.706) resulted in high confident identification. The MS^2/MS^3 approach allowed identification of phosphopeptides by spectra of relatively poor quality, which significantly improved the sensitivity for phosphorylation site mapping. It is known that phosphotyrosine (pY) containing peptides are relatively stable and often do not lose phosphoric acid to form predominant neutral loss peaks.[31] Thus, no MS^3 spectra are available for these phosphopeptides, and so a limitation of the MS^2/MS^3 approach is that peptides, which are phosphorylated only at tyrosine site, would not be identified in this strategy. Because of no neutral loss for these peptides, their MS^2 are more likely to be of good quality and could be easily identified only by MS^2.

The sensitivity of this method was also investigated by analyzing different amount of α-casein with and without Ti^{4+}-IMAC enrichment of phosphopeptides after tryptic digestion (Table S6 of Supporting Information 1). It was found that when the amount of tryptic α-casein decreased from 5 μg to 10 ng, the number of the localized phosphorylation sites decreased from 18 to 6 and 10 to 6 with and without enrichment, respectively. It is clear that this method is very powerful for phosphorylation mapping, even when the individual phosphoproteins are at the nanogram level. It should be mentioned that the prior enrichment step is very effective for phosphorylation site mapping when the individual phosphoprotein is at the microgram level; however, the enrichment step can be skipped when the individual proteins are at the nanogram level. This may be resulted from the sample loss during the phosphopeptide enrichment process. For example, when phosphoproteins

(29) Stensballe, A.; Andersen, S.; Jensen, O. N. *Proteomics* **2001**, *1*, 207–222.
(30) Sweet, S. M. M.; Creese, A. J.; Cooper, H. J. *Anal. Chem.* **2006**, *78*, 7563–7569.
(31) Bodenmiller, B.; Mueller, L. N.; Mueller, M.; Domon, B.; Aebersold, R. *Nat. Methods* **2007**, *4*, 231–237.

Scheme 1. Schematic Diagram of the MS^2/MS^3 Target-Decoy Strategy Combined with the Multiprotease Digestion Approach for Phosphorylation Site Mapping



decreased to 100 ng, the recovery of quadruply phosphorylated peptide (NTMEHVpSpSpSEEpSIISQETYKQEK) of tryptic α-casein may be much lower due to its strong interaction with the Ti^{4+}-IMAC adsorbents, which leads to the missing identification of the four phosphorylation sites (S23, S24, S25, and S28 of α-S2-casein) by adopting the phosphopeptide enrichment procedures. However, the quadruply phosphorylated peptide was successfully identified for analysis of a 100 ng sample without prior phosphopeptide enrichment. Thus, higher sensitivity for mapping of phosphorylation sites of individual proteins may be achieved by directly analyzing phosphoprotein digests, when only a minute individual phosphoprotein sample is available.

Multiprotease Digestion Approach. Comprehensive mapping of phosphorylation sites in individual phosphoproteins requires obtaining as complete sequence coverage as possible. Adoption of multiple proteases for digestion of target phosphoproteins is a practical way to improve protein sequence coverage and phosphorylation site coverage for phosphorylation site analysis of individual proteins.[14,32] In order to increase phosphorylation site coverage, sequence-specific proteases and low-specificity proteases were also used for protein digestion in this work (Table 1). The scheme of the multiprotease approach combined with the MS^2/MS^3 target-decoy strategy is outlined in Scheme 1. The phosphoprotein sample was separately digested with multiple proteases. Phosphopeptides were then separately enriched by Ti^{4+}-IMAC from individual peptide mixtures followed by LC-MS^2-MS^3 analysis. The acquired MS^2 and MS^3 spectra were then processed by the MS^2/MS^3 target-decoy strategy. The localized phosphorylation sites by each protease are outlined in Tables 2 and 3. We have identified a total of 21 of 25 phosphorylation sites of α-casein (7 of 10 phosphorylation sites of α-S1-casein and 14 of 15 phosphorylation sites of α-S2-casein) and 7 of 8 phosphorylation sites of β-casein (Tables S1 and S2 in Supporting Information 1 for the identified phosphorylation sites and their corresponding phosphopeptides; refer to Supporting Information 2 for the MS^2 and MS^3 spectra of the identified unique phosphopeptides). If the phosphoproteins were digested with a single sequence-specific protease trypsin, 16 phosphorylation sites from α-casein were identified, and only two phosphorylation sites were identified from β-casein. Especially, the four phosphorylated sites (S30, S32, S33 and S34) on β-casein could not be identified by trypsin digestion; however, they were successfully identified by proteinase K and thermolysin digestion. The four measurements for the overall performance of phosphorylation site mapping using the multiproteases digestion approach for α-casein were *Sn* (84.00%), *Sp* (96.15%), *Ac* (92.21%), and *MCC* (82.00%), which was better than the using only trypsin (Table 4). A similar result was obtained from tryptic β-casein as shown in Table 4. It is obvious that more comprehensive phosphorylation site maps could be obtained by using multiple proteases digestion.

Above results clearly demonstrated that multiple protease digestion coupled with the MS^2/MS^3 strategy could improve the sensitivity for phosphorylation site mapping. However, besides positive identifications (the known sites), there were also some false positive identifications (not reported previously) achieved for α-casein and β-casein. In order to predict the possible phosphorylation sites on α-casein and β-casein, the computational software of GPS 2.0 (http://bioinformatics.lcd-ustc.org/gps2/), which is a useful tool for predicting protein phosphorylation sites and their cognate protein kinases (PKs), was used.[25] It was found that the two novel phosphorylation sites of α-casein and three novel phosphorylation sites of β-casein localized in this study (false positive identifications) were matched with the predicted sites in highest stringency level (Table 5). This indicated that these false positive identifications are not necessarily inaccurate.

The computation time for database searching for the identification of phosphopeptides is much longer than that for the identification of unmodified peptides due to the setting of multiple dynamic modifications. The database search time will be further increased when nonspecific enzymes are used for digestion of proteins as much more peptides will be generated in silico. Herein,

(32) 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. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 7900–7905.

Table 6. Average Computation Time for One MS Spectrum Spent on Searching against Proteome and Targeted Databases with Different Enzymes[a]

	proteome database[b]		targeted database[c]	
enzyme name	MS^2 (s)	MS^3 (s)	MS^2 (s)	MS^3 (s)
Glu-C	2.34	3.51	0.10	0.13
trypsin	3.55	6.01	0.12	0.17
elastase	9.43	13.16	0.12	0.16
thermolysin	8.54	14.77	0.14	0.26
proteinase K	114.45	465.90	2.48	10.38

[a] The same 1000 MS spectra were used to investigate the computation time spent on searching against proteome and targeted databases with different enzymes on the same computer (see the cleavage sites of the enzymes in Table 1). [b] Proteome database includes bovine database (32947 entries) and its reversed version. [c] Targeted database includes 4 target proteins and 1000 decoy proteins.

the average computation time for one MS spectrum spent on searching against proteome database and targeted databases with different enzymes was investigated (Table 6). As shown, when nonspecific enzyme proteinase K was used, the computation time for global phosphoproteome analysis sharply increased to 114.45 s (32 times longer than that using trypsin) for average one MS^2 spectrum and 465.90 s (77 times longer than that using trypsin) for average one MS^3 spectrum. In a single 100 min LC-MS^2-MS^3 analysis, about 10000 MS^2 spectra and more than 4000 MS^3 spectra were acquired, that means it will cost us about 34 days to search against bovine database (totally 65894 entries). Therefore, it is not feasible for global phosphoproteome analysis for such a long time on database search. This is why most part of phosphoproteome analysis is performed by avoiding usage of nonspecific enzyme. The situation is different for the phosphorylation analysis of individual proteins when proteinase K was used. The average computation time spent on one spectrum is 2.48 s for MS^2 and 10.38 s for MS^3. Compared with global phosphoproteome analysis, the computation time is dramatically decreased. This is because that the target protein is known, the corresponding database can be much smaller. Hence, using a nonspecific enzyme is feasible for the mapping of phosphorylation sites on individual phosphoproteins in terms of computation time.

Effect of Background Proteins on Phosphorylation Site Mapping. The aim of this study is to present an approach to comprehensively map phosphorylation sites on individual phosphoproteins. Generally speaking, phosphoproteins to be analyzed are typically purified from very complex protein mixtures, and the purification is often not very specific. Thus, presence of some background proteins with the phosphoproteins is often unavoidable. These phosphopeptides derived from background proteins may interfer with localization of phosphorylation sites on the phosphoproteins of interest. To investigate this influence, we conducted phosphorylation site mapping of the cyclic AMP-dependent protein kinase (PKA) sample, which was purchased from Sigma (product number P5511). As the PKA was extracted from bovine heart, some background proteins may also have been presented in the sample. To identify these proteins, the PKA sample was digested by multiple proteases separately, and the resultant digests were analyzed by nano-LC-MS^2. The acquired MS^2 spectra were then searched using Sequest against a composite database, including the original bovine database and a reversed version of the bovine database. Finally, 261 proteins were identified from this sample at FDR ≤ 2% (see Table S3 in Supporting Information 1 for the complete list of identified proteins and their peptides). Among the identified proteins, four PKA subunits, i.e., type I-alpha regulatory subunit (IPI00714984, SWISS-PROT: P00514), type II-alpha regulatory subunit (IPI00693176, SWISS-PROT: P00515), catalytic subunit alpha (IPI00696203, SWISS-PROT: P00517), and catalytic subunit beta (IPI00693602, SWISS-PROT: P05131−1), were identified. Though this sample has PKA activity, it is far from pure. Many abundant proteins coexist with PKA subunits. To avoid the interference of phosphopeptides derived from background proteins on the identification of phosphopeptides derived from phosphoproteins of interest during database searching, we included all of these proteins in the composite database for phosphorylation site mapping. Therefore, a composite database, including all 261 identified proteins and 1000 reversed yeast sequences, was constructed for phosphorylation site mapping of the PKA sample. The procedure for mapping the phosphorylation site of PKA using the multiple protease digestion approach coupled with the MS^2/MS^3 strategy was the same as that for α-casein and β-casein. Finally, 64 phosphorylated proteins were identified by controlling the confidence of phosphopeptide identification with FDR ≤ 2% (see Table S4 in Supporting Information 1 for the identified phosphoproteins and phosphopeptides). The four PKA subunits were also found to be phosphorylated (see Table S5 in Supporting Information 1 for the details of identified phosphopeptides from four PKA subunits; refer to Supporting Information 2 for the MS^2 and MS^3 spectra of identified unique phosphopeptides from PKA subunits). As shown in Table 7, a total of 17 phosphorylated sites were identified from the 4 proteins, including 5 novel phosphorylated sites and 12 known sites. The above results indicated that the combination of the multiple protease digestion approach and the MS^2/MS^3 strategy is able to comprehensively map phosphoproteins of interest, even in presence of some background proteins.

In the above case, proteins presented in the sample were first identified, and then a composite database including these proteins was constructed for phosphorylation site mapping. This two-step approach was very time-consuming and labor intensive. In most cases, the sequences of interested phosphoproteins are known, and background proteins presented in the sample are unknown. If the inclusion of background proteins in the composite database has no significant effect on the performance of phosphorylation site mapping on interested phosphoproteins, then the first step could be skipped. To investigate this possibility, we applied a composite database containing only the sequences of the 4 PKA subunits and 1000 reversed yeast proteins for a database search. The localized phosphorylation sites are also listed in Table 7. The phosphorylation sites identified by trypsin and proteinase K were the same for both databases, while two phosphorylation sites failed to be identified by thermolysin when the small database was used. It is nice that one of the two phosphorylation sites could be identified by proteinase K. Thus, overall only one phosphorylation site failed to be identified when the database containing only the PKA subunits and yeast decoy sequences was used. On the basis of the number of the identified phosphorylation sites, we can conclude that the sensitivity of phosphorylation mapping was not

Table 7. Phosphorylation Sites of a Cyclic AMP-Dependent Protein Kinase (PKA) Sample Identified by the MS^2/MS^3 Target-Decoy Strategy Combined with the Multiprotease Digestion Approach

		MS^2/MS^{3a}				
PKA subunit		trypsin	Glu-C	elastase	thermolysin	proteinase K
type I-alpha regulatory subunit	S76[b]	√♦				√♦
	S82[b]	√♦				√♦
	S100[b,c]					
type II-alpha regulatory subunit	S45[c]				√♦	√♦
	S48[c]				√♦	√♦
	T49[d]				♦	√♦
	S75[b,c]				√♦	√♦
	S77[b,c]				√♦	√♦
	S96[b,c]	√♦				
	S380[d]	√♦				
catalytic subunit alpha	S11[b,c]					
	S15[d]	√♦				
	S140[b,c]				♦	
	T196[b]	√♦				
	T198[b,c]	√♦				√♦
	T202[b]					
	S263[d]	√♦				√♦
	S339[b,c]	√♦			√♦	√♦
catalytic subunit beta	S325[d]	√♦			√♦	
	S342[b]				√♦	

[a] √ Database search against a composite database including 4 PKA subunits and 1000 decoy proteins. ♦ Database search against a composite database including 261 target proteins and 1000 decoy proteins. [b] Phosphorylation site information from ExPasy (http://www.expasy.org). [c] Phosphorylation site information from Phospho.ELM (http://phospho.elm.eu.org). [d] Phosphorylation sites localized in this study but not reported previously.

significantly compromised for the phosphorylation site locations on the target proteins of interest when the sequences of background proteins were not included in the composite database, and also no extra novel phosphorylation sites (most probably false positive identifications) were identified when the small database was used, which indicated that the phosphopeptides derived from other 60 phosphoproteins did not lead to false positive identifications of phosphopeptide from PKA proteins. This is largely attributed to the highly confident identification of phosphopeptides by the MS^2/MS^3 approach. Thereby, the confidence of phosphorylation site mapping was not compromised by using a small database either. The above results confirmed that the inclusion of background proteins in the composite database has no significant effect on the performance of phosphorylation site mapping on the phosphoproteins of interest, and so the identification of background proteins before phosphorylation site mapping of the proteins of interest was not necessary for most cases. The exception in these cases may be that the phosphoproteins of interest coexisted with many highly abundant phosphoproteins. Then, identification of these abundant proteins for inclusion in the composite database is necessary.

In order to investigate the reliability of identified phosphorylation sites, we used the computational software of GPS 2.0 to predict the phosphorylation sites of PKA, and the results are listed in Table 8. In the five novel phosphorylation sites localized in this study, four of them were matched with the predicted sites in the highest stringency level, and one was matched with the predicted ones in the medium stringency level. So, these novel phosphorylation sites may be true positive identifications. PKA is a key enzyme in the modulation of intracellular processes in eukaryotes and is also implicated in several human diseases.[33−35] The predicted kinases responsible for the phosphorylation of the sites on PKA are also listed in the Table 8. This information may be useful for further studies of the biological function of PKA.

Table 8. GPS 2.0 Screening of New Phosphorylation Sites of PKA Subunits Localized in This Study

		GPS 2.0 prediction	
PKA	sites	threshold	kinase
type II-alpha regulatory subunit	T49[c]	high	CDK6
	S380[c]	high	PKCe
catalytic subunit alpha	S15[c]	high	PKCe
	S263[c]	medium	CAMK2a
catalytic subunit beta	S325[c]	high	PKG2

[c] Phosphorylation sites localized in this study but not reported previuosly.

CONCLUSION

It was our aim to develop a method that would facilitate comprehensive, sensitive, and reliable phosphorylation site mapping of individual phosphoproteins. To realize this goal, we presented a modified target-decoy database searching strategy for the first time to control the confidence of phosphopeptide identification for phosphorylation site analysis of individual phosphoproteins by using a much smaller composite database, including only target protein sequences and a small decoy database. Because the confidence of phosphopeptide identifications could be easily assessed by the fraction of decoy identification, no manual interpretation of spectra was required to localize phosphorylation sites. Four standard measurements of *Sn*, *Sp*, *Ac*, and *MCC* were defined to evaluate the performance of phosphorylation site mapping. As the information obtained from neutral loss MS^3

(33) ChoChung, Y. S.; Pepe, S.; Clair, T.; Budillon, A.; Nesterova, M. *Crit. Rev. Oncol. Hematol.* **1995**, *21*, 33–61.
(34) Aandahl, E. M.; Aukrust, P.; Skalhegg, B. S.; Muller, F.; Froland, S. S.; Hansson, V.; Tasken, K. *Faseb J.* **1998**, *12*, 855–862.
(35) Kammer, G. M. *Arthritis Rheum.* **1999**, *42*, 1458–1465.

and its corresponding MS^2 was combined in the MS^2/MS^3 target-decoy approach, the sensitivity and confidence for phosphorylation site analysis was significantly improved. The coverage of phosphorylation site mapping was further improved by multiple protease digestion. It has been proved that this methodology is very powerful for mapping phosphorylation sites of a sample containing one or a few individual phosphoproteins, which should be valuable for understanding various signaling events in which the phosphorylated residues in proteins partake and to further learn about the biological function of phosphoproteins and how they work.

ACKNOWLEDGMENT

Financial support is gratefully acknowledged from the National Natural Sciences Foundation of China (20675081 and 20735004), the China State Key Basic Research Program (Grants 2005CB522701 and 2007CB914102), the China High Technology Research Program (Grants 2006AA02A309 and 2008ZX10002-017), the Knowledge Innovation program of CAS (KJCX2.YW.HO9 and KSCX2-YW-R-079), and the Knowledge Innovation program of DICP to H.F. Zou, the China High Technology Research Program (Grants 2008ZX1002-020) to M. L. Ye, and from the National Natural Sciences Foundation of China (20605022 and 90713017) to M.L. Ye and (30700138) to Y. Xue.

NOTE ADDED AFTER ASAP PUBLICATION

This manuscript originally posted ASAP on June 12, 2009. The manuscript was reposted to the Web on June 16, 2009 with minor corrections to the text.

SUPPORTING INFORMATION AVAILABLE

Supplemental tables are in Supporting Information 1, and the labeled spectra of identified unique phosphopeptides are in Supporting Information 2. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review April 2, 2009. Accepted May 20, 2009.

AC900702G