Anal. Chem. **2010,** *82,* 6168–6175

Classification Filtering Strategy to Improve the Coverage and Sensitivity of Phosphoproteome Analysis

Xinning Jiang, Mingliang Ye,* Guanghui Han, Xiaoli Dong, 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

Data dependent neutral loss triggered MS3 methodology (NLMS3) is often applied to acquire MS data for the analysis of phosphopeptides. Some phosphopeptides tend to seriously lose the phosphate and result in MS2 spectra with poor fragments and fragment-rich MS3 spectra, while some phosphopeptides do not lose phosphate and result in nice MS2 spectra. Since different phosphopeptides have fragment spectra with different characteristics, filtering all of the phosphopeptide identifications by setting a global filter criteria may be inappropriate and result in low sensitivity. In this study, we developed a classification filtering strategy to improve the phosphopeptide identification and phosphorylation site localization. Phosphopeptide identifications were classified into four classes according to their different characteristics, and then, the identifications from each class of mass spectra were processed and filtered separately using different filtering strategies. It was found that the overlap of phosphopeptide identifications from different classes was low and the classification strategy significantly improved the coverage of the phosphoproteome analysis. Compared with MS2 strategy and multiple stage activation (MSA) strategy, NLMS3 with the classification filtering strategy was demonstrated to have higher sensitivity and higher performance in localizing the phosphorylation to specific sites.

Protein phosphorylation is one of the most important protein posttranslational modifications (PTM), about 30% or more proteins are phosphorylated at some point during their life cycle.[1] It plays a key role in eukaryotic signal transduction, gene regulation, and metabolic control in cells. Abnormal phosphorylation is a cause of various diseases, including cancer.[2,3] The accurate identification of phosphorylation sites and the quantification on phosphoproteins and the understanding the dynamics of this modification in response to cellular and environmental stimulus are, thus, critical for elucidating the systems biology of complex disease mechanisms and global regulatory networks.

Tandem mass spectrometry has become one of the most commonly used tools for high throughput identification and quantification of protein phosphorylation, as it cannot only identify the amino acid sequence of a peptide but also pinpoint the localization of the phosphorylation site within the sequence.[4−11] One of the challenges in protein phosphorylation study is the fact that phosphorylation is generally a labile modification and the phosphorylated peptides often produce poor fragment information in tandem mass spectrometry using collision-induced dissociation (CID). This is because the energy needed to dissociate the phosphorylation bond is much lower than that of a peptide amide bond. Thus, a large percentage of phosphopeptides undergo a significant neutral loss of phosphate, and the fragmentation of the peptide backbone yields few or no sequence ions.[11] This severely hampers efficient backbone fragmentation by MS/MS and reduces the ability of database searching algorithms to unambiguously identify the phosphopeptides.

Some bioinformatics approaches including phosphorylation specific database search algorithms and pre- or postsearch process tools have been developed to improve the sensitivity and the reliability for phosphopeptide identification and phosphorylation site localization.[12−15] These tools showed superior performances compared with commonly used database search algorithms which

* To whom correspondence should be addressed. M.Y.: address, National Chromatographic R&A Center, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China; tel, +86-411-84379620; fax, +86-411-84379620; e-mail, mingliang@dicp.ac.cn. H.Z.: address, National Chromatographic R&A Center, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China; tel, +86-411-84379610; fax, +86-411-84379620; e-mail, hanfazou@dicp.ac.cn.

(1) Cohen, P. *Trends Biochem. Sci.* **2000**, *25* (12), 596–601.
(2) Hunter, T. *Cell* **2000**, *100* (1), 113–127.
(3) Mazanetz, M. P.; Fischer, P. M. *Nat. Rev. Drug Discovery* **2007**, *6* (6), 464–479.
(4) 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* (33), 12130–12135.
(5) Gruhler, A.; Olsen, J. V.; Mohammed, S.; Mortensen, P.; Faergeman, N. J.; Mann, M.; Jensen, O. N. *Mol. Cell. Proteomics* **2005**, *4* (3), 310–327.
(6) Larsen, M. R.; Thingholm, T. E.; Jensen, O. N.; Roepstorff, P.; Jorgensen, T. J. D. *Mol. Cell. Proteomics* **2005**, *4* (7), 873–886.
(7) Lee, J.; Xu, Y. D.; Chen, Y.; Sprung, R.; Kim, S. C.; Xie, S. H.; Zhao, Y. M. *Mol. Cell. Proteomics* **2007**, *6* (4), 669–676.
(8) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. *Cell* **2006**, *127* (3), 635–648.
(9) Pinkse, M. W. H.; Uitto, P. M.; Hilhorst, M. J.; Ooms, B.; Heck, A. J. R. *Anal. Chem.* **2004**, *76* (14), 3935–3943.
(10) Trinidad, J. C.; Specht, C. G.; Thalhammer, A.; Schoepfer, R.; Burlingame, A. L. *Mol. Cell. Proteomics* **2006**, *5* (5), 914–922.
(11) Villen, J.; Beausoleil, S. A.; Gerber, S. A.; Gygi, S. P. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (5), 1488–1493.
(12) Tanner, S.; Shu, H. J.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. *Anal. Chem.* **2005**, *77* (14), 4626–4639.
(13) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. *Nat. Biotechnol.* **2006**, *24* (10), 1285–1292.
(14) Lu, B. W.; Ruse, C. I.; Yates, J. R. *J. Proteome Res.* **2008**, *7* (8), 3628–3634.
(15) Payne, S. H.; Yau, M.; Smolka, M. B.; Tanner, S.; Zhou, H. L.; Bafna, V. *J. Proteome Res.* **2008**, *7* (8), 3373–3381.

10.1021/ac100975t
Published on Web 06/22/2010

are not specifically optimized for phosphorylation analysis. However, as these approaches were still based on the MS2 spectra with poor fragments, some phosphopeptides with extreme trends of neutral loss cannot be identified. An alternative method to overcome this limitation is to trigger an additional circle of MS3 from the neutral loss peak in MS2 (data-dependent neutral loss-triggered MS3, NLMS3).[4,7,16] Because the labile phosphate group was lost in the previous stage of MS2, the MS3 spectra tend to be with more fragments. The sensitivity and reliability are significantly improved when the information of MS2 and MS3 were combined for the phosphopeptide identifications.[17−19]

A widely used approach for the automated evaluation of false discovery rate (FDR) for peptide identifications is the target−decoy strategy which is based on the principle that incorrect matches have an equal probability of being derived from either the target or the decoy database.[20−22] After database searching against a composite protein database containing both target and decoy sequences, the overall FDR of all the peptide identifications can be easily determined by setting a proper uniformed criterion. However, as the identification scores for most of the phosphopeptides only using MS2 spectra are often suppressed due to the lack of backbone fragment in MS2 spectra, sensitivity for phosphopeptide identification may be decreased by setting a uniformed criterion for all the peptide identifications, especially for the data set with a big proportion of unmodified peptides.[23] Ulintz et al. incorporated the identification probability of MS3 to generate a posterior identification probability for phosphopeptide identified from a MS2/MS3 spectra pair, which reduced the suppression for phosphopeptide identifications.[18] As the probability of unmodified peptides and phosphopeptides identified from only MS2 spectra and the probability of phosphopeptides identified by MS2/MS3 spectra were not determined in the exact same way, it may be unsuitable to use global probability criteria for the filtering of all peptide identifications. Recently, it was reported that the use of local FDR for the calculation of peptide posterior probability can improve the sensitivity for peptide identification as the peptides in the local area are with similar properties and, thus, the probability accurately reflects the actual probability of individual peptide identification.[24,25] This suggested to us that peptide identifications with different characteristics may be grouped and filtered by different criteria if the uniform criteria are hard to determine. Therefore, a classification filtering strategy was developed for phosphoproteome analysis in this study.

In NLMS3 strategy, an additional MS3 spectrum is triggered if there is significant neutral loss peak within the n highest peaks in MS2 spectra, and thus, phosphopeptides that do not lose the phosphate as the major pathway will not produce MS2/MS3 spectra pairs in mass spectrometry. The classification filtering strategy developed in this study can fully benefit from the additional fragment in MS3 and also cover the phosphopeptides that do not lose phosphate as the major pathway. After the combination of phosphopeptides identified from different classes of mass spectra, higher sensitivity was achieved for phosphopeptide identifications than that for MS2 and multiple stage activation (MSA) methodology.

METHODS AND MATERIALS

Data Sets and Database Search. The LTQ (linear trap quadrupole mass spectrometer, Thermo Finnigan) data set of phosphopeptides was acquired from Ti^{4+} affinity chromatography (Ti^{4+} IMAC) microsphere enriched human liver samples.[26] The LTQ-FTICR data set of phosphopeptides from a yeast sample was as previously reported by Ultiliz et al.[27] Each phosphopeptide sample was analyzed using three different methodologies of MS2, NLMS3, and MSA, respectively. The raw mass spectra were then preprocessed for the determining of actual charge states for MS2/MS3 spectra pairs, renewing MS3 precursor m/z values. After that, all the data sets were searched against composite databases containing both target and decoy protein sequences. Details of the data sets and the database searches were described in the Supporting Information.

Classifying the Phosphopeptide Identifications. In NLMS3 strategy, an additional MS3 spectrum is triggered only if there is significant neutral loss peak within the n highest peaks in MS2 spectra. Therefore, phosphopeptides that do not lose phosphate as the major pathway in mass spectrometry will not produce MS2/MS3 spectra pairs. In this study, the classification filtering strategy was developed to fully benefit from the additional fragment in MS3 and also cover the phosphopeptides which do not lose the phosphate as the major pathway. A flowchart of the classification filtering strategy was shown in Figure 1. There were eight steps for the identification of phosphopeptides: (1) determine neutral loss peak intensity and classify the MS2 spectra into two classes, one was a MS2 spectrum with significant neutral loss (the neutral loss peak was higher than 50% of base peak in intensity) and with consecutive MS3, and the other was MS2 spectrum without significant neutral loss; (2) perform the peak list preprocess module to determine the precursor ion charge state, remove invalid MS3 spectra, and renew the precursor ion mass of MS3 spectra if the MS1 was collected using a high accuracy mass spectrometer; (3) perform MS2 and MS3 target−decoy database searches separately; (4) combine the search results of MS2 and its corresponding MS3 and reassign the identification scores for phosphopeptide identifications from MS2/MS3 spectra pairs; (5) if there is no identical peptide match within the top 10 matched peptides of MS2 and MS3, the top matched peptides of MS2 spectra and MS3 spectra will be selected and classified as peptide identifications from NeutralMS2 spectra and NeutralMS3 spectra;

(16) DeGnore, J. P.; Qin, J. *J. Am. Soc. Mass Spectrom.* **1998**, *9* (11), 1175–1188.
(17) Jiang, X. N.; Han, G. H.; Feng, S.; Jiang, X. G.; Ye, M. L.; Yao, X. B.; Zou, H. F. *J. Proteome Res.* **2008**, *7* (4), 1640–1649.
(18) Ulintz, P. J.; Bodenmiller, B.; Andrews, P. C.; Aebersold, R.; Nesvizhskii, A. I. *Mol. Cell. Proteomics* **2008**, *7* (1), 71–87.
(19) Xu, H.; Wang, L.; Sallans, L.; Freitas, M. A. *Proteomics* **2009**, *9* (7), 1763–1770.
(20) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4* (3), 207–214.
(21) Jiang, X. N.; Jiang, X. G.; Han, G. H.; Ye, M. L.; Zou, H. F. *BMC Bioinf.* **2007**, *8*, 323.
(22) Peng, J. M.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2* (1), 43–50.
(23) Dai, J.; Jin, W. H.; Sheng, Q. H.; Shieh, C. H.; Wu, J. R.; Zeng, R. *J. Proteome Res.* **2007**, *6* (1), 250–62.
(24) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. *J. Proteome Res.* **2008**, *7* (1), 40–44.
(25) Jiang, X. N.; Dong, X. L.; Ye, M. L.; Zou, H. F. *Anal. Chem.* **2008**, *80* (23), 9326–9335.
(26) Yu, Z. Y.; Han, G. H.; Sun, S. T.; Jiang, X. N.; Chen, R.; Wang, F. J.; Wu, R. A.; Ye, M. L.; Zou, H. F. *Anal. Chim. Acta* **2009**, *636* (1), 34–41.
(27) Ulintz, P. J.; Yocum, A. K.; Bodenmiller, B.; Aebersold, R.; Andrews, P. C.; Nesvzhskii, A. L. *J. Proteome Res.* **2009**, *8* (2), 887–899.

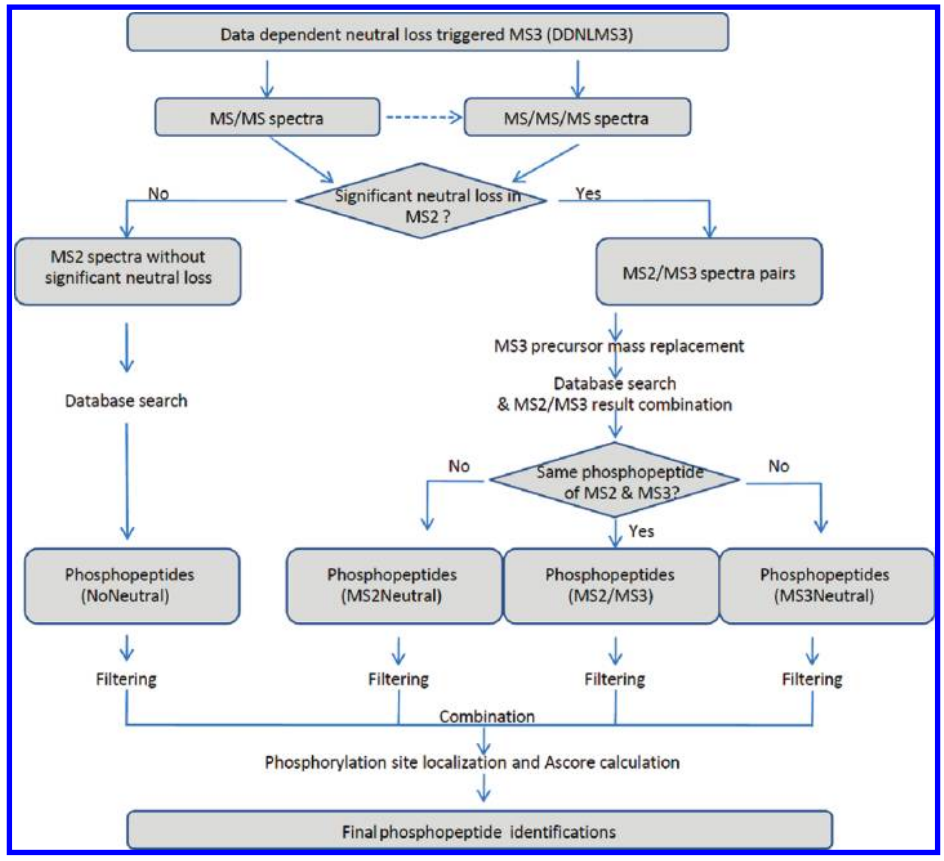

Figure 1. Flowchart of the classification filtering strategy.

(6) filter the candidate phosphopeptides from above four classes to specific FDR separately; (7) determine the most probable phosphorylation site localization and calculate the probability of the phosphorylation localized on the sites; (8) combine the phosphopeptide identifications from different classes and generate the final phosphopeptide identifications.

In the first step, all the collected spectra were separated into two groups based on the intensity of neutral loss peak in MS2. The group without significant neutral loss (neutral loss peak is less than 50% of base peak in intensity) or without consecutive MS3 spectra was termed as NoNeutral class. After that, the group of mass spectra with significant neutral loss was further separated into three classes based on if the MS2/MS3 pair could generate the same peptide assignment after database searching. If the MS2/MS3 pair can generate the same peptide assignment, the identified phosphopeptides and their spectra were termed as MS2/MS3 class. If not, the phosphopeptide assignments of MS2 and MS3 were termed as NeutralMS2 and NeutralMS3 classes, respectively. Finally, the mass spectra and phosphopeptide identifications were grouped into four classes. To control the confidence, the initial phosphopeptide identifications were filtered to FDR ≤ 1% for each class with different criteria. The filtering of phosphopeptide identifications for NoNeutral, NeutralMS2, and NeutralMS3 classes was quite straightforward, as each peptide was identified by only one fragment spectrum. Similar to a conventional target−decoy approach, phosphopeptide identifications were filtered by Xcorr and ΔCn scores (SEQUEST) or ion-score (Mascot) for these classes. However, for the MS2/MS3 class, each of the phosphopeptide identifications was derived from both MS2 and MS3 spectra. To combine the scores of MS2 and MS3 for each phosphopeptide, new scores were defined, as described in the following section. The effectiveness of the use of the newly defined scores to filter the phosphopeptide identifications achieved by a SEQUEST database search to specific FDR was demonstrated in our previous work.[17]

Defining New Scores for Phosphopeptide Identifications Derived from MS2/MS3 Spectra Pairs. The scheme of the MS2/MS3 target−decoy strategies for the SEQUEST database search algorithm was as reported previously.[17] Within the top 10 scored peptide identifications, only phosphopeptides, which were identified from both MS2 and MS3 in the spectra pairs, were retained and the newly defined identification scores were calculated by combining the match scores of MS2 and MS3 spectra; then, the top matched phosphopeptide with the highest new score for each MS2/MS3 spectra pair was kept as the most probable identification.

In our previous study, the score combination was developed for the SEQUEST database search algorithm.[17] In this study, the score combination algorithm of MS2/MS3 strategy for another most popular database search algorithm, Mascot, was designed. Similar to SEQUEST, we defined a new combined identification score, ion-score′, to combine the match information from MS2 and MS3 spectra. The ion-score′ can be calculated as

$$\text{ion-score}' = \text{ion-score (MS2)} + \text{ion-score (MS3)}$$

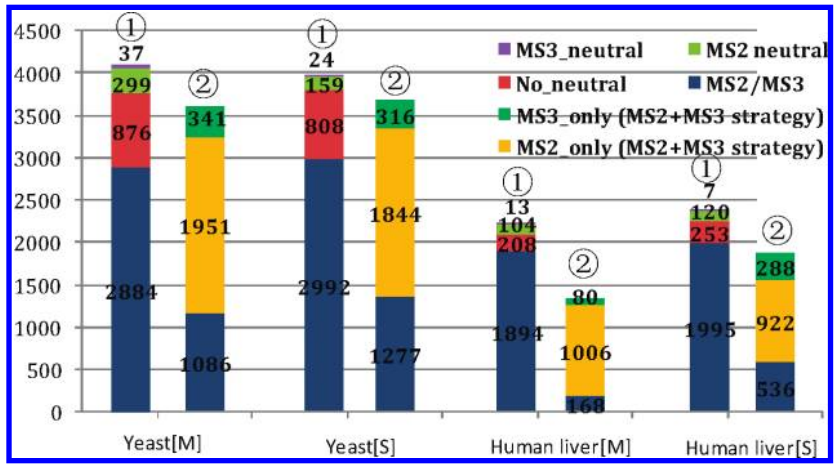

Figure 2. Cumulate number of phosphopeptides identified by different types of mass spectra with FDR less than 1% for classification filtering strategy (series ①) and MS2 + MS3 strategy (series ②). The letter in each of the brackets indicates the database search algorithm: M, Mascot; S, SEQUEST.

where ion-score (MS2) and ion-score (MS3) were the ion-scores for the same phosphopeptide identified by MS2 and MS3, respectively.

Filtering of Phosphopeptides Identified from Both MS2 and MS3 Spectra. Then, phosphopeptide identifications from MS2/MS3 spectra pairs with specific FDR can be generated by setting proper filters using the newly defined scores. For SEQUEST, the filters were similar as described previously: $\Delta Cn'_m$ filter was set as 0.1 for all charge states; then, the $Xcorr'_s$ filter was adjusted so that FDR of the identified peptides was less than a specific value (e.g., FDR ≤ 1%).[17] Mascot peptide identifications were filtered directly by ion-score′ to a specific confidence level.

Filtering of Phosphopeptides Identified from Only MS2 Spectra or Only MS3 Spectra. For Mascot database search results, ion-score was used as filter. While for SEQUEST search results, ΔCn filter was set as 0.1 for all the charge states, and then, the Xcorr filter was set for each charge state separately so that FDR of the identified phosphopeptides for each charge state was less than a specific value. For peptides identified from NeutralMS2 or NeutralMS3 spectra, filters were set separately to generate phosphopeptides with specific FDR. After the generation of phosphopeptide identifications from NeutralMS2 and NeutralMS3 spectra, if both NeutralMS2 and its consecutive NeutralMS3 spectra can generate high confident identifications, the unmatched phosphopeptide identifications from MS2 and MS3 spectra will be removed and not considered for final identification.

Localizing the Phosphorylation Sites. For peptides with multiple possible phosphorylation sites, a custom version of Ascore algorithm[13] was used to determine the most probable phosphorylation site localization. For phosphopeptides identified from NoNeutral, NeutralMS2, or NeutralMS3 spectra, MS2 (or MS3) spectra were used for the localization of phosphorylation sites. While for phosphopeptides identified from MS2/MS3 spectra pairs, in order to fully benefit from the fragment-rich MS3 spectra, both MS2 and MS3 spectra were used for the phosphorylation site localization. First, all the possible candidate permutations (MS2 candidates) for the identified phosphopeptides with different phosphorylation site localizations were generated and calculated for the Peptide Score using MS2 spectra. As MS3 spectra were acquired from the phosphopeptide ions after the loss of phosphate, MS3 candidates, which contain a phosphorylation site of dehydrate (−18, loss of phosphate), were determined for each of the MS2 candidates (e.g., for MS2 candidate of SpENSpEEDTVCR, there are two MS3 candidates, SpENSnEEDTVCR and SnENSpEEDTVCR [n indicates the dehydration while p is the phosphorylation]). Then, the peptide score for each of the MS3 candidates was calculated using the MS3 spectrum. Only the MS3 candidate with the biggest peptide score for each of the MS2 candidates was retained. After that, the sum of the peptide score of the MS2 candidate and its corresponding biggest scored MS3 candidate was used to determine the most probable phosphorylation sites. Then, Ascores were calculated for the MS2 candidate and MS3 candidate separately against the MS2 spectrum and the MS3 spectrum, respectively. The final Ascore was then calculated as the maximum value of MS2 Ascore and MS3 Ascore.

Availability of Software. APIVASEII which implemented the classification filtering strategy is integrated in ArMone, a software suite targeted for the validation and processing of the phosphoproteome data set.[28] ArMone is free for academic usage without any limitation; it can be downloaded from http://bioanalysis.dicp.ac.cn/proteomics/software/ArMone.html.

RESULTS AND DISCUSSION

Performance of the Classification Filtering Strategy. In this study, a classification filtering strategy was developed by classifying and filtering the phosphopeptide identifications separately according to the different characteristics of mass spectra. First, the performance of this strategy on the phosphopeptide identifications was investigated. Shown in Figure 2 were the numbers of phosphopeptide identifications from different classes of mass spectra using different database search algorithms (FDR ≤ 1%). As can be seen, phosphopeptides identified from the MS2/MS3 class composed the major parts of total phosphopeptide identifications. For the yeast data set, when the database search algorithm was SEQUEST (S), the classification filtering strategy identified 3983 phosphopeptides, in which 2992 (75.1%) were identified by MS2/MS3 spectra pairs; while for Mascot (M), the percentage of phosphopeptides identified from MS2/MS3 spectra was 70.4%

(28) Jiang, X.; Ye, M.; Cheng, K.; Zou, H. *J. Proteome Res.* **2010**, *9* (5), 2743–2751.

Table 1. Number of Identified Distinct Phosphopeptides from Different Classes of Mass Spectra and the Totally Identified Phosphopeptides for the Classification Filtering Strategy[a]

	MS2/MS3 pairs	NoNeutral	NeutralMS2	NeutralMS3	total	% increase
yeast (M)	1403	542	241	34	1875	33.6%
yeast (S)	1403	501	140	16	1785	27.2%
human liver (M)	563	89	76	10	671	19.2%
human liver (S)	637	126	92	6	762	19.6%

[a] The letter in each of the parentheses indicates the database search algorithm: M, Mascot; S, SEQUEST. The improvement of phosphopeptide identification by the classification filtering strategy compared to the MS2/MS3 strategy is also indicated (% increase).

(2884/4096). For the human liver data set acquired by LTQ, the percentages of phosphopeptides identified by MS2/MS3 spectra pairs were 85.5% for Mascot and 84% for SEQUEST. Most of the phosphopeptides were identified from MS2/MS3 spectra pairs, especially for mass spectra acquired by a low accuracy mass spectrometer, indicating that most of the phosphopeptides lose the phosphate as the major pathway in CID.

For phosphopeptides identified by only MS2 spectra or MS3 spectra (NoNeutral, NeutralMS2, and NeutralMS3), MS2 spectra without significant neutral loss (NoNeutral) generated the biggest number of phosphopeptide identifications. The phosphopeptides identified from this type of mass spectra accounted for 65−88% of the total number of phosphopeptide identifications, and the NeutralMS3 spectra identified the fewest phosphopeptides. For MS2 and MS3 spectra pairs which do not match to the same phosphopeptides, NeutralMS2 spectra identified many more phosphopeptides than those by NeutralMS3 spectra. This may be because the low quality of MS3 spectra significantly hampers the identification of phosphopeptides from MS3 spectra.

For comparison, the numbers of phosphopeptide identifications from the same NLMS3 data sets using conventional strategy (MS2 + MS3 strategy) were also shown in Figure 2. In MS2 + MS3 strategy, MS2 and MS3 spectra were directly submitted to database searching and data processing without classification. Peptide matches from MS2 and MS3 mass spectra were separately filtered, and then, the identifications from MS2 and MS3 were combined as the final identifications. As can be seen, the classification filtering strategy led to more phosphopeptide identifications for all the data sets acquired using either low or high accuracy mass spectrometers. Interestingly, among all phosphopeptide identifications using the MS2 + MS3 strategy, the proportion of that identified from both MS2 and consecutive MS3 spectra was significantly smaller than that from the classification filtering strategy (Figure 2). In the MS2 + MS3 strategy, the phosphopeptide identifications from MS2 and MS3 spectra were filtered before combination of the results which removed all phosphopeptide identification with low match scores. While in the classification strategy, the identification scores from both MS2 and MS3 were combined before filtering which allowed phosphopeptide identifications with relative poor matching scores to be kept. This is the main reason that more phosphopeptides could be identified by MS2 and MS3 in the classification approach. In addition, even though the classification filtering strategy generated more phosphopeptide identifications than the conventional MS2 + MS3 strategy, fewer nonphosphorylated peptides were identified by the classification filtering strategy, indicating the high performance of the classification filtering strategy for phosphopeptide identifications (Supplemental Table 1, Supporting Information).

Compared with the MS2/MS3 strategy for which only phosphopeptides identified from both MS2 and MS3 were retained for final filtering,[17] by incorporating phosphopeptides identified from only MS2 spectra and MS3 spectra, the classification filtering strategy significantly improved the number of phosphopeptide identifications (Table 1). When the database search algorithm was Mascot, the number of phosphopeptides identified from the yeast data set by MS2/MS3 spectra pairs was 1348; while considering phosphopeptide identifications from only MS2 spectra and MS3 spectra, the number of phosphopeptide identifications by the classification filtering strategy increased to 1875, 33.6% more than the MS2/MS3 strategy. For other data sets and database algorithms, the improvement of phosphopeptide identifications by the classification filtering strategy was from 19% to 33%.

In the classification filtering strategy, the target−decoy strategy,[20–22,25] one of the most commonly used strategies for FDR determination, was used to evaluate the confidence for phosphopeptide identifications. To demonstrate whether the FDR determined in this work using the target−decoy strategy can reflect the true FDR, manual validation of the phosphopeptide identifications was performed to evaluate the actual FDR. Criteria used for manual validation were similar as described before:[17,29] briefly, MS2 and/or MS3 spectra should be of good quality and with more than three sequential b- or y-type fragment ions; for phosphopeptides identified from MS2/MS3 spectra or NeutralMS3 spectra, dominant neutral loss peaks should be seen in MS2 spectra. Finally, 500 of the phosphopeptides identified from human liver sample (Mascot) and 500 from yeast sample (Mascot) were randomly selected and validated manually. For the human liver data set, 493 phosphopeptides were validated as true positives; the actual FDR was 1.4%. While for the yeast data set, four phosphopeptides (FDR = 0.8%) were considered as false positives. The above results suggested that the predicted FDR by the target−decoy strategy and the classification filtering strategy could well reflect the actual FDR. All of the spectra validated in this study can be found at http://bioanalysis.dicp.ac.cn/proteomics/Publications/APIVASEII/manual_val.htm.

Characteristics of Different Classes of Mass Spectra and Phosphopeptide Identifications. The average MS2 matching scores for phosphopeptide identification of NoNeutral, NeutralMS2, and MS2/MS3 and the average MS3 matching scores

(29) Yang, F.; Stenoien, D. L.; Strittmatter, E. F.; Wang, J. H.; Ding, L. H.; Lipton, M. S.; Monroe, M. E.; Nicora, C. D.; Gristenko, M. A.; Tang, K. Q.; Fang, R. H.; Adkins, J. N.; Camp, D. G.; Chen, D. J.; Smith, R. D. *J. Proteome Res.* **2006**, *5* (5), 1252–1260.

Table 2. Average Identification Scores for Phosphopeptides Identified from Different Classes of Mass Spectra under the Confidence Level of FDR Less than 1%

		human liver (LTQ)				yeast (LTQ-FT)			
		NoNeutral	NeutralMS2	MS2/MS3	NeutralMS3	NoNeutral	NeutralMS2	MS2/MS3	NeutralMS3
Mascot (ion-score)		59.55	57.18	42.33 (MS2) 43.05 (MS3)	60.65	41.97	38.63	35.20 (MS2) 37.12 (MS3)	45.94
SEQUEST (Xcorr)	1+	2.01	1.87	1.59 (MS2) 1.61 (MS3)					
	2+	3.8	3.85	3.50(MS2) 2.81 (MS3)	3.9	3.69	3.44	3.55 (MS2) 2.81 (MS3)	3.26
	3+	5.25	5.15	4.49(MS2) 3.35 (MS3)	4.42	4.96	4.85	4.52 (MS2) 3.55 (MS3)	4.16
	≥4+					4.81	4.71	4.22 (MS2) 3.84 (MS3)	4.22

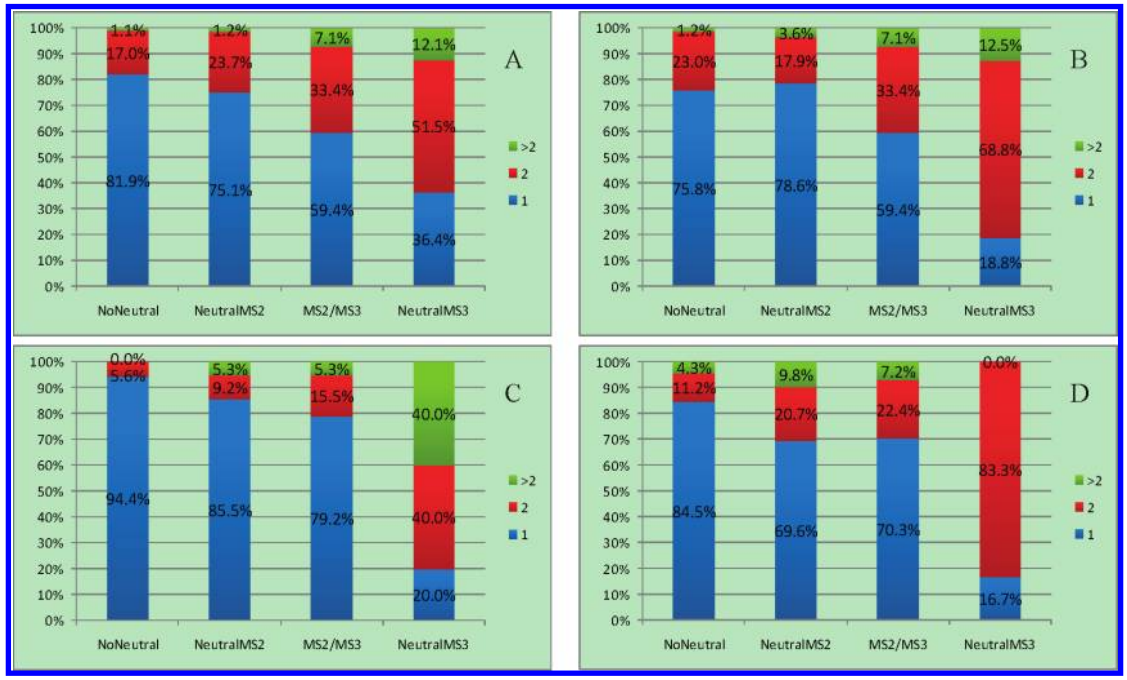

Figure 3. Percentages of the singly, doubly, and multiply phosphorylated peptides identified from different classes of mass spectra. (A) yeast (Mascot), (B) yeast (SEQUEST), (C) human liver (Mascot), and (D) human liver (SEQUEST).

for MS2/MS3 and NeutralMS3 were given in Table 2. As can be seen, phosphopeptides identified from NoNeutral mass spectra were with the highest MS2 matching score, while the MS2 matching score of MS2/MS3 class mass spectra was the lowest, and phosphopeptides of NeutralMS2 was with a matching score higher than MS2/MS3 but lower than NoNeutral mass spectra. For the MS3 matching score, phosphopeptides identified by NeutralMS3 spectra were commonly with a higher score than those identified by MS2/MS3 spectra. Evidently, phosphopeptides without significant neutral loss commonly were identified with higher scores than phosphopeptides that lose the phosphate as the major pathway. This is consistent with the fact that the significant neutral loss would hamper the identification of phosphopeptides due to the suppression of identification scores compared with nonphosphorylated peptides and phosphopeptides without significant neutral loss. The different matching scores of phosphopeptide identifications from different classes of mass spectra indicated the different fragment characteristics of different classes of mass spectra (Representative spectra of phosphopeptide identifications from different types of mass spectra were shown in supplemental Figure 1, Supporting Information).

Illustrated in Figure 3 were the percentages of singly, doubly, and multiply phosphorylated peptides identified from different classes of mass spectra under the confidence level of FDR ≤ 1%. As can be seen, the majority (75−95%) of the phosphopeptides identified from NoNeutral spectra were monophosphopeptides, indicating that the multiple phosphorylated peptides are more likely to lose the phosphate and generate predominant neutral loss peak in MS2 spectra, and most of the phosphopeptides identified from the NeutralMS3 spectra were multiple phosphorylated peptides. This may be because the multiple phosphorylated peptide prefers to lose the phosphate as the major pathway and generates a low quality MS2 spectrum which commonly does not contain sufficient fragment information for the phosphopeptide identification. The NeutralMS3 mass spectra generated the highest percentage of multiple phosphopeptides. Then, the second and third highest percentages of multiple phosphopeptide identification were from MS2/MS3 spectra pairs and NeutralMS2 spectra, respectively, and NoNeutral mass spectra generated the lowest percentage of multiple phosphorylated peptides among the four classes of mass spectra. Evidently, different types of phosphopeptides were identified by different types of mass spectra,

suggesting the high complementary of phosphopeptide identifications. Further investigations upon the overlapping of phosphopeptides identified from different classes of mass spectra also revealed the high complementary between phosphopeptides identified from different classes of mass spectra (supplemental Figure 2, Supporting Information). Therefore, by combining phosphopeptides identified from different types of mass spectra, more comprehensive coverage of phosphopeptide identifications could be achieved.

The above data clearly revealed that the classification of mass spectra and phosphopeptide identifications depends on neutral loss and the fragmentation upon the peptide backbone. For the NoNeutral and NeutralMS2 mass spectra, because the backbone was well fragmented, phosphopeptides could be identified by only MS2 mass spectra. However, the phosphopeptides identified from NeutralMS2 were with lower MS2 matching scores than NoNeutral mass spectra because of the stronger trends of neutral loss. The incorporation of NeutralMS3 can further improve coverage of phosphoproteome data, as these phosphopeptides are commonly not identifiable by only MS2 spectra. For MS2/MS3, even though the average matching scores of MS2 and MS3 were the lowest among the four classes of mass spectra, the maximum number of phosphopeptide identifications was achieved by combining the identification information from MS2 and MS3. The different characteristics of fragment mass spectra and different identification scores demonstrated that the classification strategy was reasonable and different filter criteria could be used to improve the sensitivity for phosphopeptide identifications.

Comparison of Classification Filtering Strategy with Other Approaches. To investigate the performance of the classification filtering strategy, comparisons were made upon the data set collected by NLMS3 strategy, MS2 strategy, and MSA strategy for the analysis of the same sample. Shown in Figure 4 was the number of identified phosphopeptides by these three methodologies with the confidence level of FDR ≤ 1%. When the database search algorithm was SEQUEST, for the human liver data set (Figure 4A), the numbers of identified distinct phosphopeptides by MS2 and MSA methodologies were 662 and 644, respectively, and the number of identified phosphopeptides was 762 by NLMS3 methodology using the classification filtering strategy. There were 18.3% more phosphopeptide identifications by the classification filtering strategy than the MSA strategy and 15.1% more than the MS2 strategy. For the yeast data set, the classification filtering strategy identified 17% and 11.8% more phosphopeptides than the MS2 strategy and MSA strategy, respectively. When the data sets were searched by Mascot, the increased number of phosphopeptide identifications by the classification filtering strategy than by the MSA and MS2 strategies were 3.4% and 11.6% for the yeast data set and 7% and 15.5% for the human liver data set, respectively. Evidently, by classifying the phosphopeptide identifications from mass spectra with different properties and setting filters separately, the classification filtering strategy generated more phosphopeptide identifications and showed higher sensitivity than both MS2 strategy and MSA strategy for both high and low accuracy mass spectrometry. In addition, the majority of phosphopeptide identifications (60−77%) generated by the classification filtering

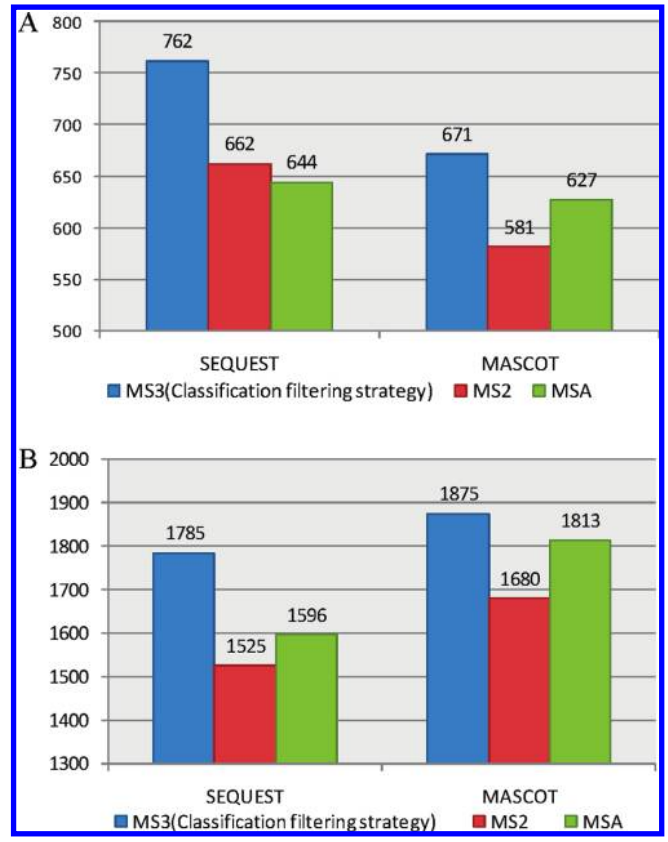

Figure 4. Numbers of distinct phosphopeptides identified from (A) human liver (LTQ) and (B) yeast (LTQ-FTICR) by MS3 (classification strategy), MS2, and MSA methods. Two database search engines were used, SEQUEST and Mascot.

strategy can also be identified by MS2 or MSA strategy, which also indicated the high confidence of phosphopeptide identifications (supplemental Figure 3, Supporting Information).

After the identification of phosphopeptides, another most important issue is to determine the localization of phosphorylation site on the peptide sequence as a big proportion of phosphopeptides are identified with more than one possible phosphorylation site localization. Ascore[13] is a very powerful strategy for the phosphorylation site localization, which calculated the probability of phosphorylation site localized on a specific site using the site-determine ions. In this study, we extended the Ascore algorithm by incorporation of MS3 spectra to improve the efficiency of phosphorylation site localization. Details of the phosphorylation site localization strategy were described in the Methods and Materials section. Performance of the phosphorylation site localization algorithm was evaluated using data sets collected by NLMS3, MS2, and MSA methods.

The percentages of the successfully localized phosphorylation sites (Ascore ≥ 19) of the identified phosphopeptides with FDR ≤ 1% for different methods were shown in Figure 5. When the database search algorithm was Mascot, there were as much as 53.6% phosphorylation sites that can be localized successfully for phosphopeptides identified by the classification filtering strategy (NLMS3) from the yeast data set, while the percentages of the successfully localized phosphorylation sites by MS2 strategy and MSA strategy were 48.9% and 47.2%. For the human liver data set, the successfully localized phosphorylation sites were 62.3%, 56.0%, and 52.5% for NLMS3 strategy, MS2 strategy, and MSA

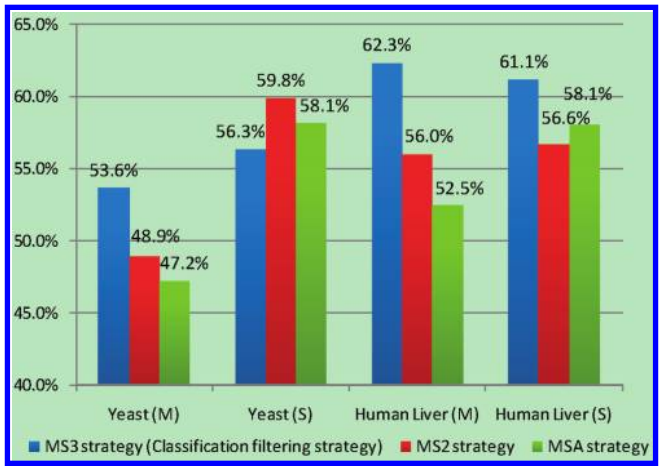

Figure 5. Percentages of the successfully localized phosphorylation sites (Ascore ≥ 19) for different data sets and database search engines. The abbreviations in the brackets indicate the database search algorithms: S is SEQUEST, and M is Mascot.

strategy, respectively. By incorporating the fragment enriched MS3 spectra, significant improvement of successfully localized phosphorylation sites were achieved compared to those obtained with the MSA strategy and MS2 strategy.

Compared with our previous work of MS2/MS3 strategy,[17] the new developed classification filtering strategy has the following major improvements: (1) classification strategy to improve phosphopeptide identification by considering phosphopeptides identified from both MS2/MS3 spectra pairs, MS2 spectra without significant neutral loss and so on; (2) handling and processing of search results from another most popular database search engine, Mascot; (3) redesigned charge state evaluation algorithm and MS3 precursor mass replacement strategy to fully benefit from the high mass accuracy; (4) custom Ascore algorithm[13] with support of the MS3 spectrum for the determination of most probable phosphorylation sites and calculating the probability of the phosphorylation localized on the sites.

In the classification filtering strategy, filtering criteria was specifically set for different classes of phosphopeptide identifications according to the different characteristics of mass spectra and different matching scores; therefore, high sensitivity should be achieved for the classification filtering strategy. Recently, Ulintz and colleagues developed another strategy by incorporating MS2 and MS3 spectra to improve the sensitivity for phosphopeptide identifications.[18] However, as the algorithm for their strategy could not be available, comparison could not be made directly. In this study, we used the yeast data set which was acquired by Ulintz et al. as described in ref 27. In their study, when the same yeast data set was used, the MSA methodology showed higher sensitivity than the NLMS3 strategy for the data processing methods of MS2/MS3_comb (combine the mass spectra before the database search).[27] There were 7% more phosphopeptides identified by the MSA strategy compared with the MS2 + MS3 strategy. In this study, using the classification filtering strategy, there were 7.6% more phosphopeptide identifications from the NLMS3 data set than that from the MSA data set when the search engine was also Mascot, and the increase of phosphopeptide identifications for the yeast data set using the SEQUEST database search algorithm was 11.8%. Significantly more distinct phosphopeptide identifications were identified by the NLMS3 data set using the classification filtering strategy than those by MS2 and MSA methodologies. The above results demonstrated that the classification filtering strategy can effectively increase the sensitivity for phosphopeptide identification using NLMS3 methodology, and the filtering of phosphopeptide identifications from MS2/MS3 spectra pairs and MS2 spectra or MS3 spectra separately showed improved discriminating performance.

The above results clearly demonstrated that the classification filtering strategy can significantly improve the sensitivity for the phosphopeptide identification using NLMS3 methodology and can effectively localize the phosphorylation sites. One issue that should be noted is that the combined scores for the filtering of phosphopeptides identified from MS2/MS3 spectra pairs may be not the optimal ones. However, the use of the simple combined scores as criteria filters did improve the sensitivity for phosphopeptide identifications. There should be additional improvement if the score combination strategies were optimized for each of the search algorithms. In addition, as there is inaccuracy of FDR evaluation by the target−decoy strategy for a small data set,[20] it may be necessary to incorporate manual validation to further improve the reliability for phosphopeptide identification, especially when the number of final identifications is not big enough.

CONCLUSIONS

A classification filtering strategy was developed to process phosphoproteome data collected by NLMS3 methodology. It was found that the different classes of mass spectra were significantly different and the resulted phosphopeptide identifications from these classes were highly complementary. By classifying mass spectra into different groups and setting special criteria for phosphopeptides identified from different classes of mass spectra, higher sensitivity and more comprehensive coverage were achieved for phosphoproteome analysis using NLMS3 methodology. It was demonstrated that the classification filtering strategy generated more phosphopeptide identifications than the MS2 and MSA methodologies, for data acquired by both a low and high accuracy mass spectrometer.

ACKNOWLEDGMENT

Financial support from the National Natural Sciences Foundation of China (20735004), the China State Key Basic Research Program Grant (2007CB914102), the China High Technology Research Program Grant (2006AA02A309), National Key Special Program on Infection diseases (2008ZX10002-017), the Analytical Method Innovation Program of MOST (2009IM031800), the Knowledge Innovation program of CAS (KJCX2.YW.HO9, KSCX2-YW-R-079) and the Knowledge Innovation program of DICP to H.Z., National Natural Sciences Foundation of China (No. 90713017), and National Key Special Program on Infection diseases (2008ZX10002-020) to M.Y. are gratefully acknowledged.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review April 13, 2010. Accepted June 9, 2010.

AC100975T