Hindawi Publishing Corporation
International Journal of Proteomics
Volume 2012, Article ID 104681, 5 pages
doi:10.1155/2012/104681

Research Article

High Mass Accuracy Phosphopeptide Identification Using Tandem Mass Spectra

Rovshan G. Sadygov

Sealy Center for Molecular Medicine, Department of Biochemistry and Molecular Biology,
The University of Texas Medical Branch, Galveston, TX 77555, USA

Correspondence should be addressed to Rovshan G. Sadygov, rgsadygo@utmb.edu

Received 27 April 2012; Accepted 18 June 2012

Academic Editor: Qiangwei Xia

Phosphoproteomics is a powerful analytical platform for identification and quantification of phosphorylated peptides and assignment of phosphorylation sites. Bioinformatics tools to identify phosphorylated peptides from their tandem mass spectra and protein sequence databases are important part of phosphoproteomics. In this work, we discuss general informatics aspects of mass-spectrometry-based phosphoproteomics. Some of the specifics of phosphopeptide identifications stem from the labile nature of phosphor groups and expanded peptide search space. Allowing for modifications of Ser, Thr, and Tyr residues exponentially increases effective database size. High mass resolution and accuracy measurements of precursor mass-to-charge ratios help to restrict the search space of candidate peptide sequences. The higher-order fragmentations of neutral loss ions enhance the fragment ion mass spectra of phosphorylated peptides. We show an example of a phosphopeptide identification where accounting for fragmentation from neutral loss species improves the identification scores in a database search algorithm by 50%.

1. Introduction

The reversible phosphorylation of proteins regulates many aspects of cell life [1–3]. Phosphorylation and dephosphorylation, catalyzed by protein kinases and protein phosphatases, can change the function of a protein, for example, increase or decrease its biological activity, stabilize it or mark it for destruction, facilitate or inhibit movement between subcellular compartments, initiate or disrupt protein-protein interactions [1]. It is estimated that 30% of all cellular proteins are phosphorylated on at least one residue [4]. Abnormal phosphorylation is now recognized as a cause or consequence of many human diseases. Several natural toxins and tumor promoters produce their effects by targeting particular protein kinases [5, 6] and phosphatases. Protein kinases catalyze the transfer of the γ-phosphate from ATP to specific amino acids in proteins; in eukaryotes, these are usually Ser, Thr, and Tyr residues.

Mass-spectrometry-based proteomics has emerged as a powerful platform for the analysis of protein phosphorylations [7]. In particular, the shotgun proteomics [8], using liquid chromatography coupled with mass spectrometry (LC-MS), has been successfully employed for comprehensive analysis of global phosphoproteome [6, 9, 10]. The advances in the phosphoproteomics were driven by developments in mass spectrometry (high resolution and mass accuracy), peptide/protein separation, phosphopeptide/protein enrichment, peptide fragmentation [11, 12], quantification, and bioinformatics data processing, Figure 1. Currently, thousands of the phosphopeptides can be detected and quantified in just one experiment. Excellent recent reviews describe experimental procedures involved in phosphoproteomics [13, 14]. Bioinformatics processing is recognized as an integral part of phosphoproteome analysis. Several applications have been developed for phosphopeptide identifications [15, 16], phosphorylation site localization [17, 18], and quantification [19]. Tandem mass spectra are searched for phosphopeptides from protein sequences with potential modifications on Ser, Thr, and Tyr residues. The searches are not targeted. Every modifiable residues can be either modified or unmodified. The effective peptide search space increases exponentially leading to computational complexity

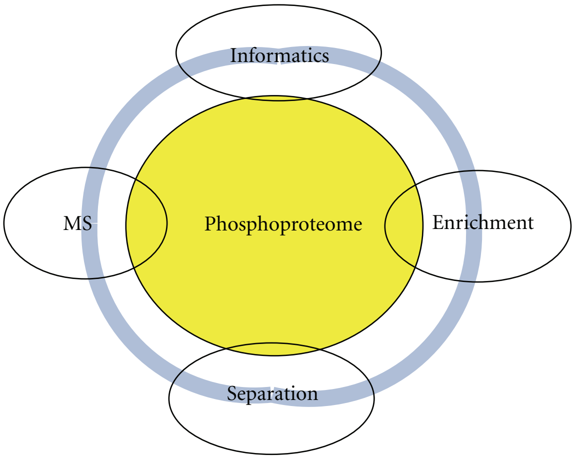

Figure 1: Phosphoproteomics and its constituent parts.

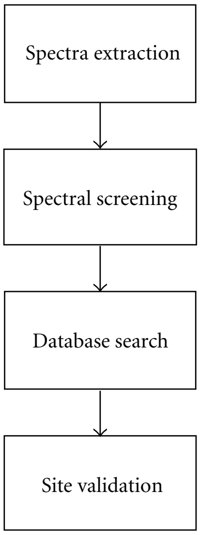

Figure 2: General informatics flowchart of a phosphoproteomics analysis.

as well as possible false identifications. High mass accuracy afforded by the modern mass spectrometers enables reducing the complexity of the search space by applying tighter bounds on peptide masses.

Lu and coworkers [20, 21] have developed models based on support vector machine (SVM) to screen for phosphopeptide spectra and validate their identifications. Their approach accurately explains spectra from phosphorylated peptides. However, SVM also acts like a black box, and it is difficult to gain insights into specifics of its decision making. Another development had used dynamic programming to relate spectra of modified and unmodified forms of a peptide [22]. This approach identifies modified peptides by comparing their tandem mass spectra with the annotated tandem mass spectra of unmodified peptides. The search space is restricted to peptides positively identified in unmodified form.

Here, we describe the informatics aspects of phosphopeptide identifications using protein sequence databases and mass spectral data from high mass accuracy and resolution instruments. Database identifications of phosphorylated peptides are done in a dynamic mode—assuming that in a peptide sequence Ser, Thr, and Tyr may or may not be are modified. For database searches, it effectively means exponential increase in the size of database. About 17% of amino acid residues (of which Ser 8.5%, Thr 5.7%, Tyr 3.0%) [23] in human proteome can potentially be phosphorylated. In general, if there are N amino acid residues which can potentially be phosphorylated, the effective database size could increase by as much as 2^N times.

2. Informatics Aspects of the Phosphoproteomics

2.1. Spectra Extraction. LTQ-Orbitrap mass spectrometer [24] stores the mass spectra in a proprietary “raw” file format (ThermoFisher Scientific, San Jose, CA). extract_msn algorithm extracts spectral information from the raw file and converts it into text file format for further data processing. It uses a built-in module to evaluate isotopic envelope of mass species. From the isotope distribution, extract_msn determines the monoisotopic mass and charge state of a peptide. Both of these are critically important and used by database search algorithms.

Normally, the full MS scan is recorded in the Orbitrap mass analyzer which is a high resolution and mass accuracy mass analyzer. The routine mass accuracy of intact peptides is in the range of ±5–10 part-per-million (ppm). This is a very high mass accuracy and is very important for reducing false discovery rates of peptide identifications. The accuracy of the intact peptide’s mass affects the number of candidate peptides from the database that will be considered in matching to the spectra. The candidate peptides are filtered based on the mass of the intact peptide and accuracy with which the mass has been measured. The higher the accuracy, the smaller the number of candidate peptides, and as a result the smaller the possibility of false positives. Fragment ion masses are recorded in ion trap mass analyzer. This is a very sensitive mass analyzer. However, the mass accuracy of measured ions is nominal, and normally in the range of ±0.5 Da. Figure 2 summarizes the informatics flowchart of a phosphoproteomics analysis.

2.2. Database Searching. Peptide identification using tandem mass spectra and protein databases is an integral part of proteomics. It is important that peptide assignments are determined with high accuracy and are verifiable. In high-throughput experiments, when thousands of tandem mass spectra are searched, it is not practical for an expert user to manually assign every spectrum and the assignments are made by software. The software uses a concept, either heuristic or probabilistic model, to measure similarity between experimental tandem mass spectrum and an amino acid sequence. For high quality spectra, when signal-to-noise ratio is high and spectra contains clearly defined ion series, most of the programs and concepts perform very well. However, when the peptide fragmentation is poor and the spectrum contains very few distinguishable peaks, or if the peptide amino acid sequence is not in the database, a number of factors lead to a wrong peptide assignment—a false identification. Depending on the software, sometimes false identifications may yield high assignment scores. High mass accuracy may substantially reduce the false identification by restricting the search space of allowed candidate sequences.

Another important aspect of the database search algorithm is the modeling of the fragmentation pattern. This is especially true for cases when there is significant difference from the routine fragmentation pattern, caused, for example, by posttranslational modifications. Thus, it is known that, in CID, peptides containing phosphorylated amino acid residues, Ser or Thr, tend to lose the phosphor group(s) before they fragment along the peptide backbone. We show here that accounting for product ions in this pathway significantly improves the identification scores of phosphopeptides in SEQUEST database search algorithm.

2.3. Assignment of Phosphorylation Sites. It is often difficult to differentiate between possible phosphorylation sites in a peptide and to uniquely assign phosphorylated amino acid residues. It has to do with the presence of several Ser, Thr, and Tyr residues in a peptide and low overall intensity of peaks from product ions of phosphorylated peptides. To address this problem, Beausoleil and coworkers [17] have developed a probability-based approach to determine phosphorylation sites of peptides from the results of SEQUEST database search algorithm [16]. Their model divides the spectra into the mass intervals of equal width. In every interval, only 6 to 8 (dependent on the intensity) peaks are retained and the rest of the peaks are ignored. In matching to the experimental peaks, only the modified fragments are considered. The probability of phosphorylation site determination is estimated via a binomial probability. This model has been successful in many practical applications especially for high scoring peptides.

3. Discussions

There are several mass spectral characteristics of phosphorylated peptides. Thus, in collision-induced dissociation reactions, one of the most prevalent pathways of the molecular dissociation is the neutral loss of the labile phosphor group of Ser or Thr residues. The presence of the relevant ion often serves as a diagnostic feature for phosphorylated peptides [25]. Often the product ions in spectra include two ion series, one from the phosphorylated peptide and the other from the precursor peptide that has lost the phosphor group(s). We have previously modified SEQUEST database search algorithm to account for the two ion sequences when identifying phosphorylated peptides [26]. The development has helped to improve the sensitivity of the phosphor peptide identifications and location of phosphorylation sites. Here, we demonstrate, in the example of this algorithm, the advantages of using extensive fragmentation pathways for targeted analysis.

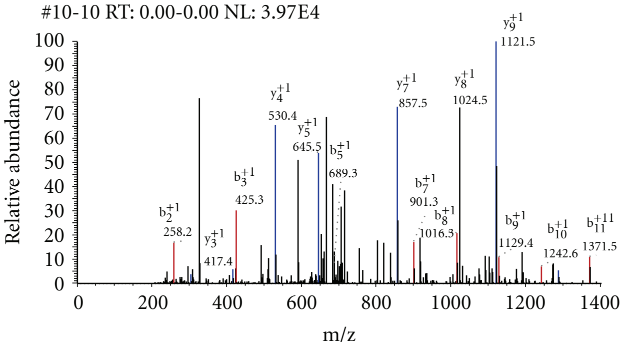

Figure 3: An example spectrum of phosphorylated peptide, R.TRS*PS*PDDILER.V. The cross-correlation scores before and after SEQUEST fragment ion modifications are 4.97 and 3.14, respectively.

In general, fragmentation of peptides via CID produces strong b- and y-ion series [27]. Therefore, most database search algorithms generate b- and y-ions (and corresponding water and ammonia losses from them) for the theoretical spectrum of a candidate amino acid sequence. In the tandem mass spectra of phosphorylated peptides, additional fragmentation patterns are observed. In addition to the b- and y-ions of the original phosphorylated peptide, ions that originate from phosphor group losses, 98 Da (from Ser or Thr residues), are also present. We accounted for these fragments by augmenting the fragmentation pattern correspondingly to add neutral loss fragments from the phosphorylated amino acid residues. The fragmentation model [28] was used for both, preliminary and cross-correlation scores in SEQUEST. Cross-correlation scores of phosphorylated peptides generated from new fragmentation pattern were about 50% higher.

The interpretations of the tandem mass spectra of phosphorylated peptides may be complicated. The main reason for this is the low abundance of fragment ions due to the alternative fragmentations. One experimental approach used for enhancing phosphopeptide identifications is to do a higher-order mass spectrometry on the fragment ions of original precursor. In these experiments, neutral loss fragment ions generated during CID in MS^2 are further dissociated generating MS^3 spectra. In spectra collected with this approach, there are number of ions corresponding to phosphoric acid group losses from b- and y-ions. An example of such a tandem mass spectrum is shown in Figure 3. SEQUEST matched this spectrum to the phosphorylated peptide sequence, R.TRS*PS*PDDILER.V. The cross-correlation and preliminary scores in the model not including phosphoric acid loss fragmentations were 3.14 and 1029.6, respectively. When we included the neutral loss ions, the cross-correlation and preliminary scores were 4.97 and 2353.6, respectively. The total number of theoretical ions generated for this peptide was 44. 35 of these ions matched to product ions in the spectrum. In contrast, 16 of the 22 theoretical ions matched the tandem mass spectrum in the original model (ignoring neutral loss fragments). The results on this and other phosphorylated peptide spectra showed that a realistic model of product ions of phosphorylated peptides needs to account for the fragments resulting from neutral (phosphoric acid group) loss of the b- and y-ions. The procedure has been automated and is used in the case of dedicated MS^n experiments to enhance fragmentation spectra of phosphopeptides.

4. Conclusion

Increased mass accuracy for precursor ions combined with enhanced fragmentation pathways helps bioinformatics methods to improve phosphopeptide identifications from tandem mass spectra and protein sequence databases. Normally identifications of phosphorylated peptides yield small cross-correlation scores. This has partially to do with the theoretical fragmentation models, which take into account only b- and y-ions generated from the peptide bond fragmentations of phosphorylated precursor peptides. We augmented the fragmentation pattern (in SEQUEST) [26] introducing theoretical peaks for b- and y-ions from neutral loss precursors and fragments. Cross-correlation scores of phosphorylated peptides increased by up to 50% using the enhanced fragmentation model.

Abbreviations

Da:	Dalton
ppm:	Parts per million
Ser:	Serine
Thr:	Threonine
Tyr:	Tyrosine
CID:	Collision induced dissociation
LC:	Liquid chromatography
MS:	Mass spectrometry.

Acknowledgments

This work was supported, in part, by UL1RR029876 UTMB CTSA (ARB), HHSN272200800048C NIAID Clinical Proteomics Center (ARB), and NIH-NLBIHHSN268201000037C NHLBI Proteomics Center for Airway Inflammation (Alex Kurosky, UTMB).

References

[1] P. Cohen, "The origins of protein phosphorylation," *Nature Cell Biology*, vol. 4, no. 5, pp. E127–E130, 2002.

[2] J. A. Ubersax and J. E. Ferrell Jr., "Mechanisms of specificity in protein phosphorylation," *Nature Reviews Molecular Cell Biology*, vol. 8, no. 7, pp. 530–541, 2007.

[3] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The protein kinase complement of the human genome," *Science*, vol. 298, no. 5600, pp. 1912–1934, 2002.

[4] P. Cohen, "The regulation of protein function by multisite phosphorylation—a 25 year update," *Trends in Biochemical Sciences*, vol. 25, no. 12, pp. 596–601, 2000.

[5] J. N. Andersen, S. Sathyanarayanan, A. Di Bacco et al., "Pathway-based identification of biomarkers for targeted therapeutics: personalized oncology with PI3K pathway inhibitors," *Science Translational Medicine*, vol. 2, no. 43, Article ID 43ra55, 2010.

[6] A. Moritz, Y. Li, A. Guo et al., "Akt-RSK-S6 kinase signaling networks activated by oncogenic receptor tyrosine kinases," *Science Signaling*, vol. 3, no. 136, article ra64, 2010.

[7] H. Zhou, J. D. Watts, and R. Aebersold, "A systematic approach to the analysis of protein phosphorylation," *Nature Biotechnology*, vol. 19, no. 4, pp. 375–378, 2001.

[8] A. J. Link, J. Eng, D. M. Schieltz et al., "Direct analysis of protein complexes using mass spectrometry," *Nature Biotechnology*, vol. 17, no. 7, pp. 676–682, 1999.

[9] E. L. Huttlin, M. P. Jedrychowski, J. E. Elias et al., "A tissue-specific atlas of mouse protein phosphorylation and expression," *Cell*, vol. 143, no. 7, pp. 1174–1189, 2010.

[10] B. Zhai, S. A. Beausoleil, J. Mintseris, and S. P. Gygi, "Phosphoproteome analysis of Drosophila melanogaster embryos," *Journal of Proteome Research*, vol. 7, no. 4, pp. 1675–1682, 2008.

[11] J. J. Coon, B. Ueberheide, J. E. P. Syka et al., "Protein identification using sequential ion/ion reactions and tandem mass spectrometry," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9463–9468, 2005.

[12] M. P. Jedrychowski, E. L. Huttlin, W. Haas, M. E. Sowa, R. Rad, and S. P. Gygi, "Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics," *Molecular & Cellular Proteomics*, vol. 10, no. 12, article M111, 2011.

[13] F. Wang, C. Song, K. Cheng, X. Jiang, M. Ye, and H. Zou, "Perspectives of comprehensive phosphoproteome analysis using shotgun strategy," *Analytical Chemistry*, vol. 83, no. 21, pp. 8078–8085, 2011.

[14] C. L. Nilsson, "Advances in quantitative phosphoproteomics," *Analytical Chemistry*, vol. 84, no. 2, pp. 735–746, 2012.

[15] B. E. Ruttenberg, T. Pisitkun, M. A. Knepper, and J. D. Hoffert, "PhosphoScore: an open-source phosphorylation site assignment Tool for MS^n data," *Journal of Proteome Research*, vol. 7, no. 7, pp. 3054–3059, 2008.

[16] J. K. Eng, A. L. McCormack, and J. R. Yates III, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *Journal of the American Society for Mass Spectrometry*, vol. 5, no. 11, pp. 976–989, 1994.

[17] S. A. Beausoleil, J. Villén, S. A. Gerber, J. Rush, and S. P. Gygi, "A probability-based approach for high-throughput protein phosphorylation analysis and site localization," *Nature Biotechnology*, vol. 24, no. 10, pp. 1285–1292, 2006.

[18] T. Taus, T. Kocher, P. Pichler et al., "Universal and confident phosphorylation site localization using phosphoRS," *Journal of Proteome Research*, vol. 10, no. 12, pp. 5354–5362, 2011.

[19] J. Cox, I. Matic, M. Hilger et al., "A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics," *Nature protocols*, vol. 4, no. 5, pp. 698–705, 2009.

[20] B. Lu, C. Ruse, T. Xu, S. K. Park, and J. Yates III, "Automatic validation of phosphopeptide identifications from tandem mass spectra," *Analytical Chemistry*, vol. 79, no. 4, pp. 1301–1310, 2007.

[21] B. Lu, C. I. Ruse, and J. R. Yates III, "Colander: a probability-based support vector machine algorithm for automatic screening for CID spectra of phosphopeptides prior to database

search," *Journal of Proteome Research*, vol. 7, no. 8, pp. 3628–3634, 2008.

[22] D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P. A. Pevzner, "Identification of post-translational modifications by blind search of mass spectra," *Nature Biotechnology*, vol. 23, no. 12, pp. 1562–1567, 2005.

[23] N. Echols, P. Harrison, S. Balasubramanian et al., "Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes," *Nucleic Acids Research*, vol. 30, no. 11, pp. 2515–2523, 2002.

[24] Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman, and R. G. Cooks, "The Orbitrap: a new mass spectrometer," *Journal of Mass Spectrometry*, vol. 40, no. 4, pp. 430–443, 2005.

[25] B. Lu, D. B. McClatchy, Y. K. Jin, and J. R. Yates III, "Strategies for shotgun identification of integral membrane proteins by tandem mass spectrometry," *Proteomics*, vol. 8, no. 19, pp. 3947–3955, 2008.

[26] R. G. Sadygov, J. Shofstahl, and A. Humer, "Improvements to the database search algorithm SEQUEST for accurate mass support and improved phosphorylation searching," in *Proceedings of the Annual Conference on Mass Spectrometry and Allied Topics*, 2005.

[27] B. Paizs and S. Suhai, "Fragmentation pathways of protonated peptides," *Mass Spectrometry Reviews*, vol. 24, no. 4, pp. 508–548, 2005.

[28] R. G. Sadygov, F. M. Maroto, and A. F. R. Hühmer, "ChromAlign: a two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces," *Analytical Chemistry*, vol. 78, no. 24, pp. 8207–8217, 2006.