Analytica Chimica Acta xxx (2017) 1–10

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Accurate phosphorylation site localization using phospho-brackets

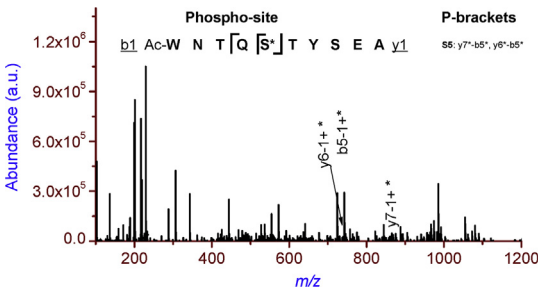
Kaijie Xiao [a, b], Yun Shen [a, b], Shasha Li [a, b], Zhixin Tian [a, b, *]

[a] School of Chemical Science & Engineering, Tongji University, Shanghai 200092, China
[b] Shanghai Key Laboratory of Chemical Assessment and Sustainability, Tongji University, Shanghai 200092, China

HIGHLIGHTS

- P-bracket was developed for phosphosite localization using paired phospho-containing site-determining product ions.
- In phosphosite localization of 101,520 synthetic phosphopeptides, a false localization rate (FLR) of 0.9% was obtained.
- 1,601 and 1,393 HeLa phosphopeptides (identified by Mascot and Sequest, respectively) were accurately localized.
- P-bracket, as a stand-alone software withuser-friendly GUIs, is available through http://proteingoggle.tongji.edu.cn.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 23 May 2017
Received in revised form
27 September 2017
Accepted 28 September 2017
Available online xxx

Keywords:
Phosphorylation
Site localization
P-bracket
Site-determining ions
False localization rate(FLR)

ABSTRACT

Phosphorylation is one of the most important and widely studied protein post-translational modifications. Tandem mass spectrometry using higher-energy collisional dissociation has evolved into a state-of-the-art analytical platform for both phosphorylation identification and site localization. Tens of thousands of phosphopeptides can now be routinely identified from a single shotgun proteomics study; site localization, however, is much more complicated and many challenges still exist. Here, we report our development of P-bracket using direct experimental evidence of phospho-containing site-determining product ions for accurate site localization without the need for additional FLR control. A P-bracket is defined as a complementary product ion pair that forms a bracket to confine a phosphorylation event to a unique site. P-bracket has been successfully benchmarked with a set of six synthetic phosphopeptides with a single phosphorylation event, a set of 96 synthetic peptides and phosphopeptide reference libraries, and two HeLa phosphopeptide LC-MS/MS (HCD) datasets; Accurate phosphosite localization by P-bracket will greatly enhance identification confidence of phosphopeptides and contribute to structural and functional studies of phosphoproteins.

1. Introduction

Phosphorylation, one of the most common and widely studied protein post-translational modifications (PTMs), plays significant regulatory roles in many cellular processes, including transduction

* Corresponding author. School of Chemical Science & Engineering, Tongji University, Shanghai 200092, China.
E-mail address: zhixintian@tongji.edu.cn (Z. Tian).

https://doi.org/10.1016/j.aca.2017.09.043
0003-2670/

of extracellular signals, intracellular transport, metabolic maintenance, and cell division [1–4]. Various aspects of phosphorylation have been widely studied by a variety of different methods [5–13]. With high accuracy, sensitivity and throughput, tandem mass spectrometry coupled with various dissociation methods has evolved into a state-of-the-art analytical platform for phosphorylation identification and site localization [14–16]; currently, phosphopeptides are often fragmented with higher-energy collisional dissociation (HCD) for ease of implementation.

With the advances in enrichment, separation, mass spectrometry and bioinformatics, tens of thousands of phosphopeptides can now be routinely identified in a shotgun proteomics study [17–21]; site localization, however, is much more complicated. On the basis of false discovery rate (FDR) control for identification, several search engines have developed search engine difference (SED) scores for further site localization of phosphopeptide candidates with the same amino acid sequence but different putative sites. The assumption is that the top hit has a significantly higher identification score than the 2nd-ranked hit and thus offers more accurate site localization as well. Representative SED scores include the Mascot Delta Score [22], SLIP score (Protein Prospector) [23], PepArML [24], and VML score (Spectrum Mill, Agilent). Besides integrated SED scores inside search engines for phosphopeptide identification, a variety of post-search tools, cataloged as probability-based localizers (PBLs) have also been developed for more specific and accurate site localization. Most of the PBLs attempt to re-assess the probability of candidate phosphopeptides with different putative sites using the "n" (normally $\leq$10) most abundant peaks per certain Th width in the MS/MS spectra. The rationale for selecting the most intense peaks is that they are less likely to be noise; however, truly informative peaks may be inadvertently missed, and all masses are not equally likely to be observed at random. Common PBLs are phosphoRS [25], Ascore [26], Phosphoscan [27], PLS (Inspect) [28], PTM Score (MaxQuant/Andromeda) [29], SLoMo [30], and Phosphinator [31]. PhosphoRS computes random match probabilities between theoretical and acquired fragment ions by applying the cumulative binomial probabilities and performing dynamic selection of peak depth. PhosphoRS reports localization scores for all of the putative sites, and a site probability $\geq$99% corresponds to a false localization rate (FLR) $\leq$1%. Ascore obtains ambiguity scores through calculation of cumulative binomial probabilities with automatic iterative selection of peak depth, and the originally proposed cutoff threshold is $\geq$ 19, which is mathematically equivalent to an FLR $\leq$1%. Essentially, both SED and PBL tools conduct site localization by re-evaluating the statistical probability of candidates with different sites using selected product ions, which may not contain phosphorylation and particular site-determining product ions. PBLs generally perform better than SED scores for site location and thus are more popular; phosphoRS followed by Ascore are the two most widely used PBL tools (Table S1).

For ambiguity assessment of site localization by different PBL tools and the establishment of a universal metric for comparison among different tools, FLR has been proposed [32]. An algorithm called LuciPHOr has been developed for the first formal FLR estimation [33,34]. LuciPHOr considers both phosphorylatable and non-native phosphorylation sites as decoys, and similar (or superior) performance with Ascore and Mascot Delta Score has been demonstrated.

Despite the general and great success of the aforementioned tools in site localization of phosphorylated peptides from their tandem mass spectra using HCD, several well-known challenges still remain: 1) appropriate addressing of intense non-sequence neutral loss (NL) [35]; 2) efficient discrimination of adjacent putative sites [22,36]; 3) peak identification in dense product ion zones; and 4) the absence of a universal decoy population for FLR estimation. Currently, FLR estimation is still substantially influenced by both database size and nature; the frequency and proximity of non-native proline and glutamic acid (serving as the decoy residues of serine and threonine) are only approximate, and different dissociation methods [37] and alternative sites of different distance [22] require different cutoff scores to reach the same level of FLR.

For accurate site localization of common PTMs, we have previously developed PTM Score as implemented in protein database search engine ProteinGoggle [38,39]. PTM Score is defined as total number of non-redundant matching product ions that independently define the unique localization of a PTM. These site-determining product ions (with or without bearing the PTM) serve as direct or indirect evidence of the site. While PTM Score works well for non-labile (i.e., no NL during dissociation) PTMs, they do not work well for labile PTMs (such as phosphorylation) because intense NL can occur. As such, the observation of product ions without phosphorylation in a phosphopeptide tandem mass spectrum does not necessarily imply that the corresponding original amino acid sequence in the peptide is not phosphorylated. In principle, non-phosphorylated product ions must not be used for site localization.

In this report, to demonstrate accurate site localization of labile phosphorylation on phosphopeptides from the corresponding tandem mass spectra using HCD, we describe the development of P-bracket using only direct experimental evidence of phospho-containing site-determining product ions. First, P-bracket was benchmarked with a set of six synthetic phosphopeptides with a single phosphorylation event; P-bracket, phosphoRS and Ascore accurately localized 100.0%, 58.3% and 33.3% of the sites, respectively. Second, the FLR performance of P-bracket was benchmarked with a set of 96 synthetic peptides and phosphopeptide reference libraries [36]; For the 141,550 phosphopeptides identified by Mascot with a global FDR of $\leq$1% from the corresponding 96 LC-MS/MS (HCD) datasets, P-bracket accurately localized 74,855 of them with a global FLR of 0.9%. Third, P-bracket was validated with a HeLa phosphopeptide LC-MS/MS (HCD) dataset [40]; P-bracket successfully localized 1601 and 1393 phosphopeptides with Mascot and Sequest identification, respectively, which outperformed both phosphoRS and Ascore.

P-bracket, in principle, can be effectively implemented into various peptide database search engines for simultaneous phosphopeptide identification and site localization without the need for an additional post-search localization tool or further specific FLR control.

2. Materials and methods

2.1. Chemical and reagents

Formic acid (FA) and methanol were HPLC grade and purchased from Sigma-Aldrich (St. Louis, MO, USA). The Millipore Simplicity system was used to produce ultra-pure water on site. Six representative synthetic phosphopeptides, which were phosphorylated on S, T, or Y (WNTQS*TYSEA, WNTQSTYS*EA, WNT*QSTYSEA, WNTQST*YSEA, AEDKTY*KYICR, and AEDKTYKY*ICR), were purchased from BankPeptide (Hefei, Anhui, China).

2.2. The six synthetic phosphopeptides

A set of six synthetic singly phosphorylated peptides with representative sites on serine, threonine and tyrosine was first used for the performance evaluation of P-bracket. Each of these phosphopeptides was individually analyzed using ESI-HCD on a Q

Please cite this article in press as: K. Xiao, et al., Accurate phosphorylation site localization using phospho-brackets, Analytica Chimica Acta (2017), https://doi.org/10.1016/j.aca.2017.09.043