Journal of Proteomics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Proteomics

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

Are neutral loss and internal product ions useful for top-down protein identification?

Kaijie Xiao, Fan Yu, Houqin Fang, Bingbing Xue, Yan Liu, Yunhui Li, Zhixin Tian *

School of Chemical Science and Engineering, Tongji University, Shanghai, China
Shanghai Key Laboratory of Chemical Assessment and Sustainability, Tongji University, Shanghai, China

ARTICLE INFO

Article history:
Received 26 October 2016
Received in revised form 13 March 2017
Accepted 15 March 2017
Available online xxxx

Keywords:
Protein identification
Neutral loss
Internal product ions
Isotopic envelope fingerprinting
Protein database search
Overlapping isotopic envelopes

ABSTRACT

Neutral loss and internal product ions have been found to be significant in both peptide and protein tandem mass spectra and they have been proposed to be included in database search and for protein identification. In addition to common canonical b/y ions in collision-based dissociation or c/z ions in electron-based dissociation, inclusion of neutral loss and internal product ions would certainly make better use of tandem mass spectra data; however, their ultimate utility for protein identification with false discovery rate control remains unclear. Here we report our proteome-level utility benchmarking of neutral loss and internal product ions with tandem mass spectra of intact *E. coli* proteome. Utility of internal product ions was further evaluated at the protein level using selected tandem mass spectra of individual *E. coli* proteins. We found that both neutral loss and internal products ions do not have direct utility for protein identification when they were used for scoring of P Score; but they do have indirect utility for provision of more canonical b/y ions when they are included in the database search and overlapping ions between different ion types are resolved.

Biological significance: Tandem mass spectrometry has evolved to be a state-of-the-art method for characterization of protein primary structures (including amino acid sequence, post-translational modifications (PTMs) as well as their site location), where full study and utilization tandem mass spectra and product ions are indispensable. This primary structure information is essential for higher order structure and eventual function study of proteins.

1. Introduction

Tandem mass spectrometry using collision-(CID/HCD [1]), electron-(ETD [2]/ECD [3,4]), and photon-(UVPD [5,6]/IRMPD [7,8]) based dissociation methods has been widely used for peptide [6,9–19] and protein [20–24] identification. The confidence of this identification is normally proportional to the total number of matching product ions, where normally only canonical b/y ions or c/z ions are used. The non-canonical neutral loss (NL) and internal product ions are also present in nearly all types of MS/MS spectra, and whether these ions should be equally adopted for protein identification has been of great interest. Some related studies have been carried out for HCD. Among 10,878 HCD spectra of doubly protonated tryptic peptides, internal product ions have been found to be up to 15%, and they were proposed to be taken into account by MS/MS search engines [25]. In a later yet larger study of >16,000 high-quality HCD spectra of tryptic peptides, Mann et al. found that average percentage of isotopic peaks of internal product ions was 13% and the corresponding percentage of abundance was 10% [26]. For intact proteins, Kelleher et al. found that each of the 76 residues in ubiquitin was covered by about 24 internal fragments during beam-type collisional dissociation [27]. In de novo peptide identification, internal product ions in HCD spectra have been used to distinguish between similar candidate peptides and improve de novo sequencing [28,29]. Both ya- and yb-internal product ions up to 30 kDa were adopted in protein database search engine Mascot-TD to be searched in addition to canonical b/y product ions and their water and ammonia loss ions; predominant internal fragmentation was observed in collision-activated dissociation of 3 standard proteins, which significantly improved sequence coverage and protein identification confidence [30]. So far, utility of either neutral loss or internal product ions for protein identification at the proteome level still remains unknown.

Here we report our initial proteome-level utility evaluation of neutral loss as well as internal product ions observed in HCD tandem mass spectra of *E. coli* proteome. Both neutral loss and internal products ions were found to have no direct utility for protein identification in terms of number of protein IDs; but they do have indirect utility for provision of more canonical b/y ions when they are included in the

* Corresponding author.
E-mail address: zhixintian@tongji.edu.cn (Z. Tian).

http://dx.doi.org/10.1016/j.jprot.2017.03.011
1874-3919/

database search and overlapping product ions between different ion types are resolved.

2. Experimental section

2.1. Materials

Formic acid (FA, eluent additive for LC-MS, 56302) and acetonitrile (CHROMASOLV gradient grade, 34851) were purchased from Sigma Aldrich (St. Louis, MO, USA). BCA Reagent Kit, NaCl, PBS, PMSF, Tryptone, and yeast extract were bought from Sangon Biotech (Shanghai, China). Millipore Simplicity system was used to produce ultra-pure water.

2.2. Cell culture of E. coli and protein extraction

The procedure for the cell culture of *E. coli* and related protein extraction has been reported in detail elsewhere [23], and only a brief description is given here. Sterilization of the conical flask with Tryptone, Yeast extract, NaCl and doubly-distilled H_2O was first done in a high-pressure steam sterilizer at 121 °C for 21 min. The flask was then disinfected in an ultra-clean bench with UV light for 15 min. *E. coli* was cultured in the flask overnight at 37 °C and 220 rpm. After centrifugation of 5 min (8000 rpm and 4 °C) and 3 times wash with 20 mL PBS each, the cell pellet was re-suspended in 5 mL PBS with 50 μL PMSF. Cells were lysed in a 1.5 mL centrifuge tube over ice using a Ningbo Scientz ultrasonic cell disruptor (5 min, 300 J, 4 °C, every cycle contains a 5 s running and a 10 s pause). After centrifugation of 15 min at 10000 rpm and 4 °C, the protein concentration in the supernatant was measured using BCA assay in TECAN Infinite F50 according to the manufacturer's protocol. The *E. coli* cell lysate was finally aliquoted into 1.5 mL centrifuge tubes and stored at a −80 °C refrigerator for future use.

2.3. RPLC-MS/MS analysis of E. coli proteome

Reversed-phase liquid chromatography (RPLC) tandem mass spectrometry using HCD of *E. coli* proteome was carried out in a Thermo Scientific Q Exactive mass spectrometer coupled with a Dionex UltiMate 3000 RSLCnano high-performance liquid chromatography (HPLC) system. The analytical column (75 μm i.d., 60 cm long) was packed in-house with C4 (5 μm, 300 Å), and the trap column was packed with the same particle but with i.d. of 200 μm and length of 5 cm. Buffer A was 5.0% ACN, 94.8% H_2O and 0.2% FA; Buffer B was 95.0% ACN, 4.8% H_2O and 0.2% FA. After loaded on the trap column, the *E. coli* proteome was eluted with the following gradient: 0 min-1% B, 1 min-15% B, 92 min-65% B, 98 min-75% B, 103 min-99% B, held at 99% B for an additional 15 min. MS spectra of the precursor ions were acquired with the following settings: microscans 2, resolution 70,000 (*m/z* 200), AGC 3E6, scan range *m/z* 600–2000. Data-dependent Top10 tandem HCD spectra acquisition settings were as following: microscans, 1; resolution, 35,000 (*m/z* 200); AGC, 5E5; maximum IT, 250 ms; isolation window, 10 *m/z*; NCE, 30%; charge exclusion, 1–4; dynamic exclusion, 20.0 s. Both MS and MS/MS spectra were acquired at the centroid mode, and three technical replicate RPLC-MSMS (HCD) datasets (D1, D2, and D3) of *E. coli* intact proteome were obtained. Three technical replicate HCD spectra (S1, S2, and S3) for each of *E. coli* proteins CSRA_ECO24, DBHB_ECO57, RL7_ECOLI, YGAU_ECOL6, and IPYR_ECOL6 were exported from each dataset in Thermo Xcalibur Qual Browser.

2.4. Database search using ProteinGoggle 2.0

Intact protein database search using ProteinGoggle 2.0 implemented with the isotopic mass-to-charge (*m/z*) ratio and envelope fingerprinting (iMEF) search algorithm was fully reported elsewhere [23,31] and only a brief description is given here.

For *E. coli* proteome, both forward and random databases containing the theoretical isotopic envelopes for both the precursor and the product ions were created from the *E. coli* flat text file. The text file, containing 7658 protein entries (2589 unique ones by amino acid sequence), is downloaded from UniProt (www.uniprot.org) with the following criteria: 'Organism [OS]' = escherichia coli, 'Sequence_Fragment' = No, 'Sequence_Sequence length' = 1–200, and 'Reviewed' = Yes. For each direction, two incremental databases with product ion series of b/y and b/y + (b/y-NL) for every charge state of each proteoform were created; and these two types of databases were designated as IS-I and IS-II, respectively. Theoretical isotopic envelopes of precursor ions were computed for all possible charge states of every proteoform in the MS acquisition window. It should be noted that a and a-NL ions (a = b-CO) were included in b-NL ions; and NL represents NH_3, H_2O, NH_3 + H_2O, $2NH_3$, or $2H_2O$ where applicable. H_2O loss comes from product ions containing amino acids D/E/S/T, and NH_3 loss comes from product ions containing amino acids K/N/Q/R. Proteoforms with combinatorial amino acid sequence and post-translational modifications (PTMs) were generated using shotgun annotation. With all annotated PTMs in the flat text file treated dynamically, 2883 proteoforms (with 294 containing PTMs) were generated in both Database I and Database II.

For selected individual proteins, forward and random databases were created for product ions. For each database, four incremental databases with product ion series of b/y, b/y + (b/y-NL), b/y + (b/y-NL) + i (i = internal) and b/y + (b/y-NL) + i + (i-NL) were created; and these four types of databases were designated as IS-I, IS-II, IS-III, and IS-IV, respectively. The numbers of theoretical product ions of each ion series for these five proteins are provided in Supplemental Table S-1.

For database search and protein identification from the data-dependent datasets (D1, D2, and D3), both precursor ion and product ion candidates were "fished" from Database I or II using isotopic *m/z* fingerprinting; and fully identified using isotopic envelope fingerprinting. Two sets of search parameters (IPACO/IPMD/IPAD) of 40/15/100 and 20/15/50 are used in this study for precursor and product ion search, respectively; IPACO, IPMD and IPAD are short names of isotopic peak abundance cutoff (%), isotopic peak *m/z* deviation (ppm) and isotopic peak abundance deviation (%), respectively. These search parameters were pre-optimized at the proteome level for most protein IDs with FDR control (data not shown). Final protein identification with a FDR of 1% at the spectrum level was achieved through decoy search using random database and P Score cutoff. The definition of P Score used in ProteinGoggle is fully described elsewhere [32]. The computation of P Score is based on Poisson distribution, and is a modified version of that used by ProSightPC [33] to accommodate the different product ion identification algorithm used by ProteinGoggle. For database search and protein identification from the individual MS/MS spectra (S1, S2, and S3), product ion candidates were "fished" from Database IS-I, IS-II, IS-III or VI using isotopic *m/z* fingerprinting and fully identified using isotopic envelope fingerprinting. The set of search parameters (IPACO/IPMD/IPAD) of 20/15/50 are used.

For the public available datasets of *Pseudomonas aeruginosa* PAO1 (http://hdl.handle.net/2022/17234) [34], the customized database containing both precursor and product ions were created in the same way with that of *E. coli*. The flat text file was downloaded from UniProt with the following criteria: 'Organism [OS]' = Pseudomonas aeruginosa, 'Sequence_Fragment' = No, 'Sequence_Sequence length' = 1–300, and 'Reviewed' = Yes. For each direction, two incremental databases with product ion series of b/y and b/y + (b/y-NL) for every charge state of each proteoform were created, respectively. With the annotated PTMs of mono-, di-, tri-methylation, phosphorylation, and acetylation in the flat text file treated dynamically, 1366 proteoforms (with 286 containing PTMs) were generated. The search parameters are the same with those of *E. coli*.

Please cite this article as: K. Xiao, et al., Are neutral loss and internal product ions useful for top-down protein identification?, J Prot (2017), http://dx.doi.org/10.1016/j.jprot.2017.03.011