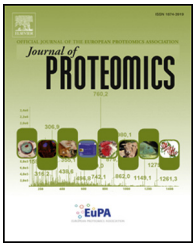
JOURNAL OF PROTEOMICS 103 (2014) 254–260

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot



Technical note

N-terminal sequence tagging using reliably determined b_2 ions: A useful approach to deconvolute tandem mass spectra of co-fragmented peptides in proteomics

*Fedor Kryuchkov, Thiago Verano-Braga, Frank Kjeldsen**

Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

ARTICLE INFO

Article history:
Received 17 December 2013
Accepted 26 March 2014
Available online 12 April 2014

Keywords:
Mass spectrometry
Complementary pairs
Mixture spectra
Proteomics
HCD

ABSTRACT

With the recent introduction of higher-energy collisional dissociation (HCD) in Orbitrap mass spectrometry, the popularity of that technique has grown tremendously in the proteomics society. HCD spectra, however, are characterized by a limited distribution of b_n-type ions, which permit the generation of reliable sequence tags based on complementary *b*,*y* pairs both for *de novo* sequencing and sequence tagging strategies. Instead, most peptide HCD spectra (~95%) are dominated with b_2 ions. In this work, we analyzed positive predictive values of b_2 ions in HCD, and found that b_2 ions can be determined with >97% certainty in the presence of a_2 and its complementary y_{n-2} ions. Analytically, b_2 ions provide information on the composition of the first two N-terminal amino acids in peptides. Their utilization in N-terminal sequence tagging leads to a significant decrease in false discovery rate by filtering out false positives while retaining true positive identifications. As a consequence, the number of peptide spectrum matches (PSMs) increased by 4.8% at fixed FDR (1%). This approach allows for deconvolution of mixture spectra and increased the number of PSM to 9.2% in a complex human sample and to 24% in a complex sample of synthetic peptides at 1% FDR.

The bottom-up approach involving tandem mass spectrometry (MS/MS) of peptides from enzymatically digested proteins dominates the field of proteomics. The central element of this approach is to derive sequence information from fragmented peptides to facilitate confident peptide identification. For the purpose of peptide identification, two major approaches are common: *de novo* sequencing and database searching. In *de novo* sequencing, peptide sequences are inferred by interrogation of the MS/MS spectrum typically through "spectrum graph" algorithms [1,2], while in database searching, peptide MS/MS spectra are scored against peptides embedded in a protein sequence database [3,4]. Merging sequence tags with database searching is termed "sequence tagging" and was first investigated by Mann et al. [5]. Later, this approach was revisited and improved by several groups [6–10]. The attractive benefits of the sequence tagging approach include significant filtering of invalid peptide candidates from the database resulting in faster

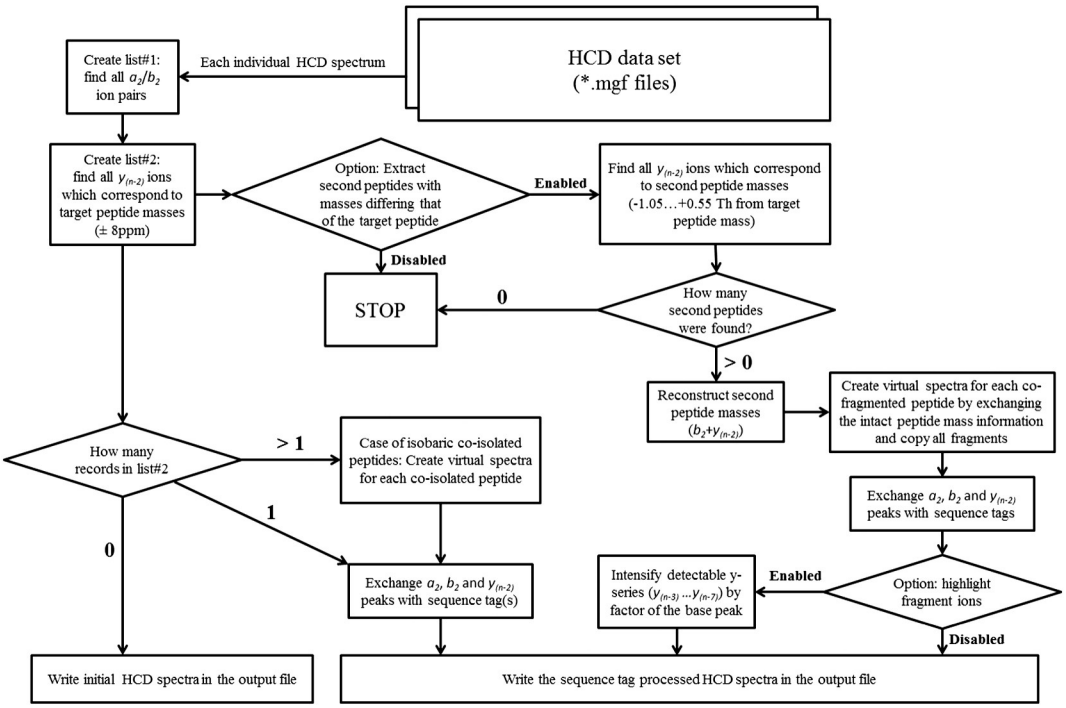
* *Corresponding author.*
E-mail address: frankk@bmb.sdu.dk (F. Kjeldsen).

http://dx.doi.org/10.1016/j.jprot.2014.03.039
1874-3919/

searches in the restricted database search space, and a higher yield and confidence of peptide identification. However, the sequence tagging approach requires caution since its success depends on the accuracy of the submitted sequence tag [11]. Advances in mass spectrometry allow for the acquisition of MS/MS spectra with high resolution (>10,000) at a frequency of more than 10 Hz, which has improved the sequence tagging technique. Still, not all peptide fragments are equally valid targets for sequence tagging. Complementary pairs formed internally in MS/MS or those formed between orthogonal fragmentation techniques have been proven to be highly reliable [8,9,12–14], and complementary pairs have been employed in several sequence tagging approaches [8,9,15,16]. Nevertheless, with the recent extension of this application to HCD in Orbitrap MS, the complementary approach is facing new challenges. While the presence of many complementary sets of *b*,*y*-ions can be found in ion-trap CAD spectra [14], the distribution of b_n-ions is significantly reduced in HCD [17,18]. This can be explained by the higher internal energy transfer of ions fragmented in beam-type analyzers *versus* those activated by slow-heating in ion-traps. Thus, larger *b* ions rapidly decay to the more stable b_2-ion [18,19]. b_2-ions do not decay to any significant degree to the shorter b_1 ion (the b_1 ions are far less thermodynamically stable), but statistical analysis by Zubarev and co-workers showed that abundant CO loss (−27.9949 Da) from b_2-ions occurs to form a_2 ions in 95% of all HCD spectra [18]. Here, we evaluate the reliability of b_2-ions and test their utility for N-terminal sequence tagging. Additionally, the potential for deconvolution of frequently generated mixture spectra in large-scale proteomics will be explored using this N-terminal sequence tagging approach.

We analyzed an HCD MS/MS dataset recorded from a tryptic digest of human LoVo cells. The cells were harvested and lysed, and the extracted proteins were reduced, alkylated, digested with trypsin, and separated as described in Supporting Information A1 prior to MS/MS analysis. Both MS and MS/MS spectra were recorded with high resolution, processed and searched as described in Supporting Information A3. HCD data were pre-processed prior to database searching using the a_2b_2-restrictor software (Scheme 1; software freely available from http://composition.sdu.dk/) that interrogates each peptide spectrum entry (Mascot Generic Format, mgf) and performs the following steps: 1) create list #1 of found a_2/b_2 pairs; 2) search list #1 for complementary y_{n-2} ions (as [M+2H]-b_2) and create list #2 of confirmed $a_2/b_2/y_{(n-2)}$ hits; and 3) if list #2 is empty — the original spectrum peak list is written without any changes to the mgf-file; if list #2 consists of one hit the software writes in a new mgf-file all the original spectrum peaks with the corresponding sequence tag in the header and omits the corresponding redundant spectrum peak information ($a_2/b_2/y_{n-2}$); if list #2 consists of more than one unique hit of ($a_2/b_2/y_{n-2}$) ions, multiple virtual spectra are written for each hit while omitting corresponding $a_2/b_2/y_{n-2}$ ions as described.

Since the reliable sequence tag of b_2 ions provides information about the first two N-terminal amino acids (AA) additional restrictive power was gained in database searching. The AA composition of b_2 ions can be deduced from the accurate mass. Indeed, for 20 AA, there are 210 pair combinations comprising 177 different masses. Since b_2 ions are the dominant *b* ions in HCD spectra, it is not possible to infer the order of the two AA in cases where b_1 ions are absent [19]. This does not prevent the



Scheme 1 Flowchart of the pre-processing of HCD data by the a_2b_2-restrictor software. –