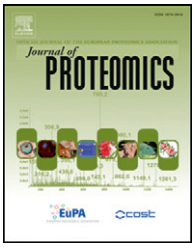
JOURNAL OF PROTEOMICS 75 (2012) 5293–5303

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot



Data extraction from proteomics raw data: An evaluation of nine tandem MS tools using a large Orbitrap data set

Francesco Mancuso[a], *Jakob Bunkenborg*[b], *Michael Wierer*[a], *Henrik Molina*[a,c,*]

[a]*Centro de Regulación Genòmica (CRG), C/ Dr. Aiguader 88, 08003 Barcelona, Spain*
[b]*Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark*
[c]*The Rockefeller University, 1230 York Ave. New York, NY 10065, USA*

ARTICLE INFO

Article history:
Received 22 March 2012
Accepted 12 June 2012
Available online 20 June 2012

Keywords:
Tandem mass spectrometry
Extraction tools
HCD
CID
Orbitrap

ABSTRACT

In shot-gun proteomics raw tandem MS data are processed with extraction tools to produce condensed peak lists that can be uploaded to database search engines. Many extraction tools are available but to our knowledge, a systematic comparison of such tools has not yet been carried out. Using raw data containing more than 400,000 tandem MS spectra acquired using an Orbitrap Velos we compared 9 tandem MS extraction tools, freely available as well as commercial. We compared the tools with respect to number of extracted MS/MS events, fragment ion information, number of matches, precursor mass accuracies and agreement in-between tools. Processing a primary data set with 9 different tandem MS extraction tools resulted in a low overlap of identified peptides. The tools differ by assigned charge states of precursors, precursor and fragment ion masses, and we show that peptides identified very confidently using one extraction tool might not be matched when using another tool. We also found a bias towards peptides of lower charge state when extracting fragment ion data from higher resolution raw data without deconvolution. Collecting and comparing the extracted data from the same raw data allow adjusting parameters and expectations and selecting the right tool for extraction of tandem MS data.

1. Introduction

The use of tandem mass spectrometry to obtain peptide sequence information is fundamental for proteomics. The speed with which mass spectrometers generate MS/MS data has improved greatly and sequencing of double digit numbers of peptides per second is resulting in primary data files exceeding gigabytes in size. Handling the vast amount of raw data requires efficient software algorithms to extract the essential data: determination of charge state and mass of the precursor and extraction of the fragmentation data. Several search algorithms that query tandem MS data have been compared and analyzed in detailed studies [1–3] but though extraction of tandem MS data has been examined [4,5] we are not aware of studies dedicated to an comprehensive comparison of the tools that process the primary data prior to database searching.

Mathematical manipulations and filters are applied in the process of converting raw data into more information dense short lists of masses and intensities but with any filter there is a risk of losing important information. Processing of primary data can involve: i) charge state determination of the precursor ion, ii) calculation of the precursor ion mass, iii) calculation of fragment ion masses, iv) charge state deconvolution of fragment ions, v) deisotoping of fragment ions and vi) general noise reduction. Because database search results depend on setting the optimal parameters based on the input data, it is important to know the pros and cons for different extraction tools.

* *Corresponding author at:* The Rockefeller University, 1230 York Ave. New York, NY 10065, USA. Fax: +1 212 327 8620.
E-mail address: henrik.molina@gmail.com (H. Molina).

1874-3919/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jprot.2012.06.012

Awareness of potential differences between tandem MS extraction tools can be used for optimizing search settings but knowledge of differences is also valuable for inter-laboratory comparisons and protocols as exemplified by ABRF initiated studies and the PRIME-XS consortium (www.primexs.eu).

The diversity of different mass spectrometers used to generate tandem MS data in proteomics is great and the same holds true for algorithms used to query the tandem MS data. For our analysis we chose tandem MS data from an Orbitrap mass spectrometer [6] and chose MASCOT as search algorithm, because this combination is of interest not only to our group but is a used combination by many other proteomics laboratories (ABRF presentation ABRF-PRG2011: "The Interaction Between Users and Suppliers of Proteomics Services/Facilities," San Antonio, TX February 19–22, 2011). We based our analysis on both low resolution and higher resolution tandem MS spectra. The lower resolution data were generated and measured in a linear ion trap (CID) and the higher resolution tandem MS spectra were generated in a quadrupole type collision cell (HCD) and measured in an Orbitrap. For Orbitrap mass spectrometers several extraction tools are available and in this study we have included 9 such tools. Six are provided by academic research groups: DeconMSn (v. 2.2.2.2) [7], DTASuperCharge (v. 2.0b1) [8], MaxQuant (v.1.0.13.13 [9,10] and v. 1.1.1.14 [11]), Raw2MSM (v. 1_10_2007_06_14) [12], and VEMS (v. 5.20092010) [13,14] and three are commercial or vendor based: Distiller (v. 2.3.2.0, Matrix Science), Extract_MSn (v. 5, Thermo Scientific) and Proteome Discoverer (v. 1.2.0.208, Thermo Scientific). Table 1 summarizes the 9 tools. Because tandem MS extraction is a process that not always receives much attention and because it is our experience that habit is often the deciding factor for the selection of tools we decided to include different versions of the same tool. This also allows us to compare incremental changes for the same software.

For this analysis, peak lists originating from the same primary data but generated by 9 tandem MS extraction tools were queried against the same protein database using MASCOT [15] and peptide matches fulfilling a set score threshold were compared between peak lists and MS/MS experiments. The analysis involves comparisons of i) number of extracted MS/MS spectra, ii) data density of extracted spectra, iii) number of matched spectra fulfilling a set score threshold, iv) mass accuracy of the extracted precursor masses, v) overlap of matched spectra in-between the tools, vi) similarity between the tools and vii) the number of unique peptides and proteins identified using the peak lists generated by each of the 9 different tools.

2. Experimental section

2.1. Generation of samples

A large set of tryptic peptides was generated using several sources including phospho peptide enrichment. The peptides were separated by reversed phase liquid chromatography and analyzed by an Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA). Peptides were fragmented either in the linear ion trap and also measured here (CID) or in a "higher collision energy" cell (HCD) and measured in the orbitrap. Detailed information is available in "Supplementary Experimental Section."

2.2. Generation of peak lists

Each .RAW file was processed by 9 tandem MS extraction tools of both commercial (Distiller, Extract_MSn and Proteome Discoverer) and academic origin (DeconMSn, DTASuperCharge, MaxQuant v. 1.0 and v. 1.1, Raw2MSM, and VEMS). The DTASuperCharge processing was done using the built-in DTA generator (not the Extract_MSn post-processing that is another option for generating peak lists). Replicated analyses were merged and submitted to MASCOT v. 2.3.01 using MASCOT Daemon 2.3. Tandem MS extraction parameters used for the extractions were chosen as default for all tools, except for Distiller where the "spectra collapsing" option was disabled. The generation of peak lists by MaxQuant v. 1.1 is very much integrated with Andromeda (MaxQuant's search algorithms). We therefore followed the standard work flow for MaxQuant v. 1.1 to generate the peak lists (named .apl) which we made MASCOT-compatible using an in-house made Perl script. Extraction parameters used for each of the 9 algorithms are listed in Supplementary Table 1 and Supplementary File 1 contains examples of the peak list format created by each of the tools. Some of the tools only allow for the adjustment of a few settings whereas others, exemplified by Distiller, offer a wealth of settings. We have used default settings for all extraction tools with the caveat that another combination of the many options might change the results.

2.3. Database searching

The generated peak lists were queried against a database using MASCOT and results were parsed using standard filters and criteria. A mass tolerance window of 10 ppm was used for precursor ions. For fragment ions mass tolerances were set to either 20 mTh (HCD) or 0.5 Th/0.8 Th (CID). Additional details are available in the "Supplementary Experimental Section." All data, extracted peak lists and search results are available at PRIDE [16] with the accession numbers: 24258 through 24352.

3. Results and discussion

3.1. Data content changes on the transformation from raw data to peak list

To gauge the efficiency in transforming raw tandem MS data into peak lists we began our comparison by calculating the percent of triggered MS/MS spectra (271,787 CID and 175,915 HCD MS/MS experiments) written to the peak lists by each of 9 different tandem MS extraction tools. For the CID data close to all MS/MS experiments were written to the peak lists (98%–100%) by each tool. However, the peak lists generated by DeconMSn and Raw2MSM contained more spectra than actually triggered MS/MS experiments (8% and 35% additional spectra, respectively). For the HCD data we observed a similar pattern: DeconMSn and Raw2MSM peak lists contain more