Sequence analysis

XDIA: improving on the label-free data-independent analysis

Paulo C. Carvalho[1,2,*], Xuemei Han[2], Tao Xu[2], Daniel Cociorva[2],
Maria da Gloria Carvalho[3], Valmir C. Barbosa[1] and John R. Yates III[2]

[1]Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Caixa Postal 68511, 21941-972 Rio de Janeiro, Brazil, [2]Department of Chemical Physiology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA and [3]Biophysics Institute Carlos Chagas Filho, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

ABSTRACT

Summary: XDIA is a computational strategy for analyzing multiplexed spectra acquired using electron transfer dissociation and collision-activated dissociation; it significantly increases identified spectra (∼250%) and unique peptides (∼30%) when compared with the data-dependent ETCaD analysis on middle-down, single-phase shotgun proteomic analysis. Increasing identified spectra and peptides improves quantitation statistics confidence and protein coverage, respectively.

Availability: The software and data produced in this work are freely available for academic use at http://fields.scripps.edu/XDIA

Contact: paulo@pcarvalho.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Label-free approaches (e.g. spectral counting) for the relative quantitation of complex peptide mixtures have gained popularity because of their low cost, reasonable accuracy and simplicity. However, in accuracy they still fall behind the approaches that measure abundance by comparing peptides to an internal, chemically identical standard enriched with a heavy stable isotope. The latter, in turn, are expensive and laborious. But regardless of the strategy, quantitation and identification, efficiency and accuracy are constrained by limitations from the data acquisition method.

The widely adopted data-dependent acquisition (DDA) relies on a full-range survey scan (MS1) to select peptide ions for fragmentation. Peptides are then identified by submitting their tandem mass spectra (MS2) to a protein database search engine such as SEQUEST or Mascot (Perkins *et al.*, 1999). In general, DDA selects the most abundant peaks in the MS1 for individual fragmentation, and the peaks for which MS2 is acquired are excluded for a short period of time from further MS2 acquisition. If there are many co-eluting ions, however, there may be insufficient time to acquire MS2 data for all ion species and thus some low-abundance ions may never be identified. Poor sampling of low-abundance ions affects overall sampling statistics, thereby impacting the accuracy of quantitation, especially for the lowest abundance peptides.

To overcome such limitations, Venable *et al.* introduced an alternative strategy to DDA that does not rely on MS1 data, called data-independent acquisition (DIA) (Venable *et al.*, 2004). Briefly, the authors used a fast scanning linear ion trap (LTQ) (Thermo Fisher, San Jose, CA, USA) to sequentially isolate and fragment precursor windows of $10\,m/z$ by collision-activated dissociation (CAD) until a desired range was covered (e.g. $400–1000\,m/z$). Their results showed a 3- to 5-fold improvement in the signal-to-noise ratio of the ion chromatograms in comparison to DDA, clearly demonstrating the benefits of DIA for quantitation (Venable *et al.*, 2004). Moreover, the method provided time-consistent ion sampling and was able to identify peptides undetected in MS1. The increased sensitivity is due to the ability of the linear ion trap to accumulate selected precursor ions for MS2, thereby becoming less affected by chemical noise than MS1. Venable *et al.* did not attempt to extract information from the multiplexed spectra that may have occurred when multiple precursor ions were fragmented in the same m/z window; this greatly compromised the number of identifications.

Here we describe a novel approach for shotgun proteomic data acquisition, termed extended data-independent acquisition (XDIA), which greatly improves on the latter and reduces the gap between the label-based and label-free technologies. We demonstrate XDIA on a middle-down proteomic experiment, which targets the analysis of bigger molecules than the ones produced by a trypsin digest. Some key advantages of analyzing large molecules include increased identification coverage and the possibility of assessing relationships among multiple modifications in the same molecule (e.g. histone). To accomplish such, our experimental strategy uses electron transfer dissociation (ETD), a new fragmentation process that is effective for larger molecules and is known to conserve the information of post-translational modifications. To enable for protein identification, the data must be processed by our software, termed the XDIA processor.

2 DATA ACQUISITION

XDIA combines high-resolution Orbitrap survey scans with multiplexed fragmentation data acquired using ETD and CAD. An XDIA data acquisition cycle consists of one high-resolution MS1 (e.g. acquired with the Orbitrap) and a series of consecutive MS2 events of $20\,m/z$ isolation windows overlapping each other by $1\,m/z$. Ion dissociation is achieved by ETD (without supplemental activation) followed by CAD in two scan events (ETD-CAD). This process is repeated until a desired m/z range is covered. Then the XDIA processor is used to enable for protein identification.

A schematic of this process is shown in Supplementary Figure 1 and the experimental method is detailed in Supplementary Method 1.

We acquired spectral counts on a yeast lysate digested with Lys-C using XDIA and compared the result to the one obtained by DDA ETCaD (ETD with supplemental activation). These analyses were performed in triplicates.

3 ALGORITHM

The generated data were processed by the XDIA processor, whose algorithm is as follows. For every MS2 isolation window, a list of monoisotopic peaks (henceforth referred to as the precursor list) having m/z values within the isolation window m/z bounds is extracted from the two nearest MS1 events (Supplementary Fig. 1). This is accomplished using subroutines from YADA (Carvalho *et al.*, 2009), a software developed by our group that is an isotopic envelope pattern recognition algorithm. This precursor list is used to remove charge-reduced precursors [CRPs, which have been proven to play a critical role for charge prediction and precursor detection in ETD MS2 (Sadygov *et al.*, 2008)] and neutral loss peaks from the corresponding ETD MS2 spectrum. Then any remaining CRP patterns are used to infer precursors that were undetected in the flanking MS1 events. These newly detected (MS2) precursors are added to the precursor list, and their CRPs and neutral loss peaks are also removed from the MS2. The final output is generated by combining information from the MS1 precursors, the MS2-detected precursors and the cleaned list of MS2 product ions. These data are then re-organized to be saved as if they were originally acquired by DDA, so they can serve as input to a protein search engine and the analysis pipeline of choice (details in Supplementary Discussion). When acquiring data using the XDIA ETD-CAD methodology, the +2 ions fragmented using ETD are excluded by the XDIA processor because of ETD's poor efficiency in dissociating +2 ions (details in Supplementary Discussion).

4 RESULTS

Table 1 shows the average number of identified spectra from each charge state. XDIA ETD-CAD identified roughly 250% more spectra than DDA ETCaD. The average and standard deviation of the number of uniquely identified peptides for the three yeast lysate replicates acquired by XDIA ETD-CAD and DDA ETCaD were 1434 ± 210 and 1035 ± 155, respectively. A general overview showing that the great majority of the peptides achieve a greater spectral count when using XDIA ETD-CAD instead of DDA ETCaD is presented in Supplementary Figure 2. A false discovery rate of <1% was obtained for all analyses (Supplementary Method 2). A Venn diagram comparing the number of uniquely identified peptides when using DDA ETCaD, DIA ETD-CAD and XDIA ETD-CAD is presented in Supplementary Figure 3. The DIA ETD-CAD

Table 1. Average number of identified spectra for each charge state from data acquired from three experiments using XDIA ETD-CAD and DDA ETCaD

Ion charge	XDIA ETD-CAD	DDA ETCaD
+2	1206	558
+3	1823	736
+4	628	249
+5	186	39
+6	21	2
+7	11	0
Total	3874	1584

results were obtained by analyzing the DIA data without using the XDIA processor, as performed by Venable *et al.* (2004).

5 FINAL CONSIDERATIONS

Boosting the number of identified spectra significantly improves the quantitation statistics confidence; increasing identified peptides improves on coverage, making XDIA a key contribution. XDIA is recommended for experiments targeting post-translational modifications, those for which ETD is recommended. XDIA performs best when analyzing complex samples with limited chromatographic separation. Such is the case of a yeast lysate in a single chromatographic run.

The XDIA processor is coded in C# and requires a PC with Windows XP SP2 or later and .NET 3.5. It installs under the directory PatternLab for proteomics (Carvalho *et al.*, 2008).

Conflict of Interest: none declared.

REFERENCES

Carvalho,P.C. *et al*. (2008) PatternLab for proteomics: a tool for differential shotgun proteomics. *BMC Bioinformatics*, **9**, 316.

Carvalho,P.C. *et al*. (2009) YADA: A tool for taking the most out of high-resolution spectra. *Bioinformatics*, **25**, 2734–2736.

Perkins,D.N. *et al*. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Sadygov,R.G. *et al*. (2008) Charger: combination of signal processing and statistical learning algorithms for precursor charge-state determination from electron-transfer dissociation spectra. *Anal. Chem.*, **80**, 376–386.

Venable,J.D. *et al*. (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods*, **1**, 39–45.