Using state of the art data independent acquisition (DIA) methods for protein identification in complex mixtures

Yves le Blanc[1]; Stephen Tate[1]
[1]SCIEX, 71 Four Valley Drive, Concord, ON, L4K 4V8 Canada

INTRODUCTION

Data dependent acquisition (DDA) methods have been the workhorse of protein/peptide identification by mass spectrometry, However, the stochastic ion selection process creates randomness that has been discussed numerous times. Also, as proteomic matrices are so complex, there is a proportion of MS/MS spectra with co-isolated precursor ions which is very difficult to deconvolute and can create identification confidence issues.
Although DIA methods have been focused on the improvement of quantitation results. The use of DIA for untargeted peak list generation for identification was originally pioneered through the PaCIFICA [reference] method and proved that there is significant potential for improvement in identification methods. DDA technology have stayed static for a long period of time with a common isolation window and very similar logic in the parent selection criteria between all MS vendors. These 2 factors impact significantly the ability to increase the number of identified species especially when modern MS instruments are extremely sensitive and co-isolation of ions is a real issue.
Data independent acquisition (DIA) methods provide a route to deconvolute the MS/MS and generate a more robust and reproducible compound identification list. Although there have been a number of attempts to undertake this (i.e. MSPLIT-DIA or DIAUmpire) they have always suffered from the lack of ability to identify the precursor with any degree of accuracy and therefore the impact on peptide identification is large. Recent developments in SWATH acquisition methods has shown that is it possible to acquire data with smaller isolation windows and cover the mass range of interest. However, these methods are slow and not compatible with modern chromatography. Scanning SWATH acquisition operates in a disconcerted manner with an isolation window which is moving with time. This investigates the use of Scanning SWATH acquisition for protein identification investigating reproducible protein lists from samples which far improve on the reproducibility of DDA and also far increase our depth of coverage in a single proteomic sample.

MATERIALS AND METHODS

K562 Cell lysate digest (SCIEX) was reconstituted according to supplier recommendation. The stock solution was diluted to a final concentration of 10ng/µL prior to sample preparation. For EvoTips C18 loading (EvoSep, Denmark), the manufacturer supplied protocol was used as is (1). Briefly, tips were first washed, conditioned, equilibrated before loading of sample (20µL). The tips were also washed loaded with 100µL of Solvent A (water+0.1% formic acid) for preservation. This yielded 200ng of material loaded on each tips which were stored up to a week at 4ºC prior analysis.

LC separation as performed with EvoSep system (Evosep, Denmark) using the 60, 100 and 200 SPD (samples per day) workflow. This provided elution gradients of 21 min (1µL/min), 11 min (1.5µL/min) and 5 min (3.5µL/min). The EV1109 Performance (8cm x 150µm, 1.5µm – 60 and 100 SPD) and EV1107 Endurance (4cm x 150µm, 1.9µm – 200 SPD) columns were coupled with a zero-dead volume union to the micro probe (1-50µL) of OptiFlow Turbo V ion source. The 1-10µL/min electrode was used in all cases and the column oven was set to 35ºC, as recommended by column supplier. The ion source temperature was set to 150ºC with GS1 and GS2 were set to 10 and 35, respectively. Scanning SWATH acquisition was performed using a 1amu window with 20ms accumulation and Rolling-CE applied over 50amu mass range (TripleTOF 6600 system, SCIEX). A total of 10 mass ranges, covering precursor m/z region of 400 to 900, were acquired in triplicate for each of the workflow evaluated.

Data processing was performed in DIANN v 1.8 using human FASTA file downloaded from UniProt August 2021. Result files from DIANN were post processed in Python to prepare figures and generate overlap maps from the data

Number of identified peptides and proteins:

The ever-ongoing arms race in proteomics means that a longer list of identified compounds is always better, or is it? However, it is known that the ever-increasing list-omics work does not always translate to the DIA workflows which are in general orientated to quantification analysis. We compared our results with our previous instruments highlight how such a workflow can be used for peptide and protein ID.

Data Set Name	Total Run Time	Sample consumed (ug)	Number of Peptides	Number of Proteins
Median 6600 Data	90	4	11954	3064
5min GFF	50	1	43971	5116
11min GFF	110	1	60323	6220
21min GFF	210	1	77971	7043

Table 1: Peptide and proteins identify at 1%FDR from the different gradient and experiment lengths.
The differing experiments show a consistent increase in the number of peptide and proteins identifications. A more significant increase is seen with the different gas phase fractionation methods and the gradient lengths. Although increasing gradient lengths is known to increase identifications to a certain degree, it was a surprise to see a 4 time increase in the gradient added 50 % more proteins.

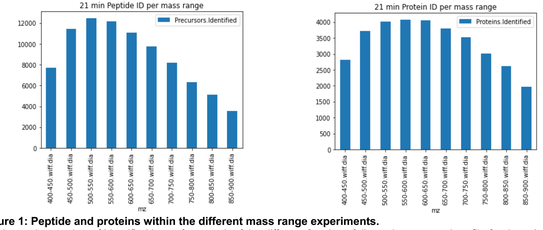

Figure 1: Peptide and proteins within the different mass range experiments.
As shown the number of identified items from each of the different fractions follows the expected profile for the m/z of a human lysate sample. The max number of identified specific is in the range of 500-550 m/z. Interestingly the profile is different for the number of proteins identified but it is long known that the length of a peptide identified has a large effect on protein identification. Although not shown it is possible to generate the same distributions for the other experiments and the picture is essentially the same.

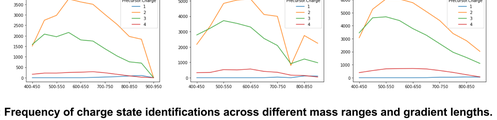

Figure 2: Frequency of charge state identifications across different mass ranges and gradient lengths.
Although the experiment used segmented mass ranges and an expectation of charge state of species in each range could be made it was interesting to see the actual profiles. What was interesting was the potential increase in the 3+ ions in the early fractions and maybe a better representation of the 4+ ions. However as somewhat expected the overall profiles are equivalent.

Reproducibility of results:

DIA methods have always promised an improvement in the data reproducibility. We investigated the repeatability from using increasing gradient length as well as using replicate experiments.

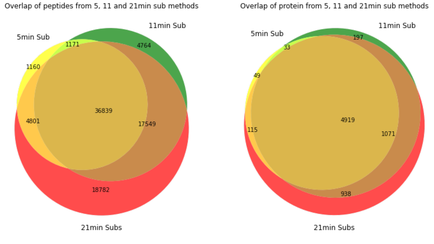

Figure 3: Overlap of identified species.
There is much made of the reproducibility of DDA methods and the ability of the stochastic engine to potentially make mistakes. DIA is supposed to be a more reproducible technique and therefore any scanning method would also hopefully show the same characteristics. This should also mean that the compounds identified in any shorter gradient method should in reality be identifiable in the longer gradient methods. The multi way Venn diagrams shown above indicate that although not a complete perfect superset the 21min data does not include 7095 peptides which were identified in shorter gradient lengths (left panel). These 7000 peptides however only equate to an extra 279 protein above the 7043 protein identified.

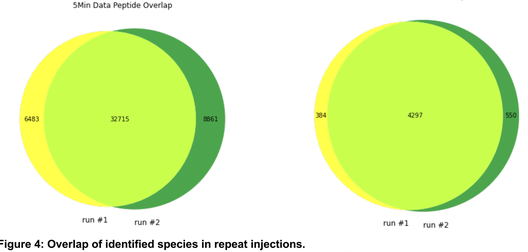

Figure 4: Overlap of identified species in repeat injections.
This figure shows the repeat analysis of the sample using the 5 min GFF method. Although it is known that the technique used for the isolation and selection of ions from the sample shows zero bias to the intensity of the precursor there is still not very reproducible results from the data extraction. Again, this is minimized through the number of proteins which are aggregates of the peptides. All things being equal the results between the different data sets should be equivalent. Although reproducibility of the longer gradient method was not undertaken it is hoped that repeat analysis using a more representative gradient will show better reproducibility.

Rate of new peptide identification increases faster than protein?

Is there a visible increase in the protein sequence coverage across the different gradient experiments.

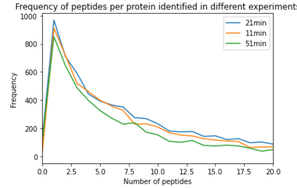

Figure 5: Unique peptides per protein identification frequency.
The large increase in the number of peptides associated with the lengthening of the gradient for each sub fraction does not appear to be associated only with a significant increase in the number of proteins as single hit wonders. The data shows that there is an increase in the number of peptides per protein but there is not a significant increase in the number of single hit wonders. So, although the protein number increase 50% from 5 mins to 21min experiments the likelihood is that there is a significant portion of the results associated with multiple peptides per protein.

CONCLUSIONS

1. Gas Phase fractionation provides superior depth of coverage of a single sample
2. Fast scanning methods are needed to provide the depth of coverage and ensure full sample is analyzed
3. For a similar given amount of sample and instrument time double the number of proteins may be possible when compared to classical DDA
4. Potentially indicates that stochastic identification rates maybe linked to a yet unknown factor

REFERENCES

1) https://www.evosep.com/wp-content/uploads/2020/03/Sample-loading-protocol.pdf

TRADEMARKS/LICENSING