Generating globally FDR-controlled DDA-based spectral libraries for analyzing wide-window DIA experiments

Jacob Lippincott[1], Susan Ludwigsen[1], Seth Just[1], Brian Searle[1,2]

[1]Proteome Software Inc., Portland, OR [2]The Ohio State University, Columbus, OH

Introduction

DIA (data independent acquisition) is rapidly becoming a popular method for label-free quantitative proteomics. Ideally, DIA searches are carried out by generating sample specific libraries from narrow-window DIA data. However, it is not always feasible to collect multiple sample-specific injections for library building. Alternatively, predicted peptide libraries may be searched, but without empirical correction this approach usually produces fewer detections and typically is limited to considering unmodified peptides (Figure 1).

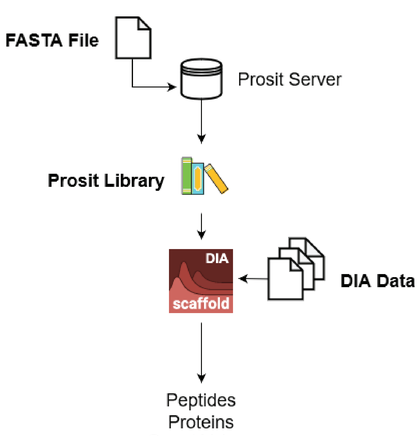

Figure 1: Prosit library workflow

Most proteomics laboratories, however, have years of existing DDA (data dependent acquisition) data that can be leveraged. Here, we present a workflow in which existing data can be used to build libraries that can be used in place of a sample-specific DIA library for a fraction of the cost and time (Figure 2)

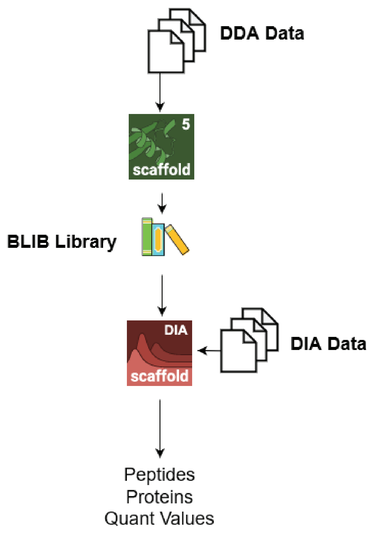

Figure 2: DDA library workflow

Experimental Methods

HeLa data from Bruderer et al. (PXD005573) [1] was accessed from the PRIDE repository. While this dataset contains iRT stable isotope labeled (SIL) peptides, these peptides were not considered as part of either library generation or for the search process. Raw fractionated DDA data was searched against a UniProt Human FASTA file using MSFragger version 3.3 in Scaffold version 5.0.1, with precursor and fragment tolerances of 4.5 and 20 ppm respectively.

Experimental Methods (Contd.)

Carbamidomethylation of C was set as a fixed modification, and oxidation of M and protein n-terminal acetylation were set as variable modifications. Peptide identifications were filtered to 1% FDR in Scaffold. The resulting peptide detections were exported in the BLIB format. Retention times in the BLIB library were corrected to account for retention time variation in the different DDA runs using the Kernel Density Alignment approach in EncyclopeDIA (Figure 3).

Yeast data from Searle et al. (MSV000084000)[2] was accessed from Massive. This data set did not contain any SIL peptides.

DDA Library Approach

Replicate wide-window DIA HeLa injections from the study were searched using Scaffold DIA version 3.1.0 against the exported library. Wide-window yeast DIA data was searched against an existing BLIB library

Ten ppm was used as precursor and fragment tolerances. Carbamidomethylation of C was specified as a fixed modification and oxidation of M was specified as a variable modification. Peptide and protein level FDR was thresholded to 1%.

Predicted Library Approach

Prosit-predicted libraries were generated using the publicly available Prosit server, with the 2019 prediction model. A UniProt Human FASTA file and a UniProt Yeast FASTA file were used.
Prosit predictions were converted to DLIB (a spectrum library format) using EncyclopeDIA version 0.9.0. The HeLa and yeast injections were searched against their respective libraries using the same precursor and fragment tolerances mentioned above.

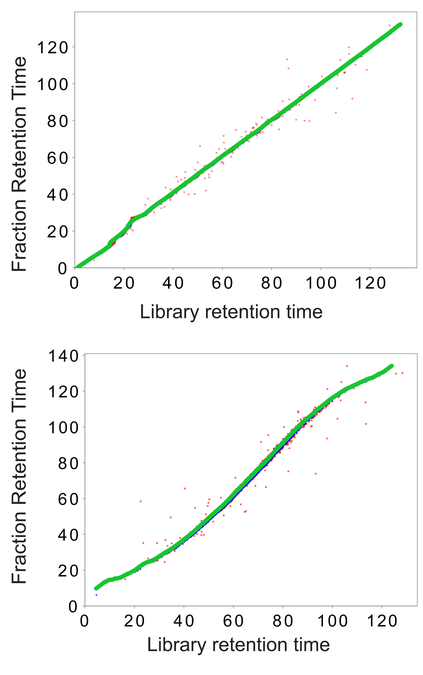

Figure 3: Retention time alignment curves produced via Kernel Density approach. RT alignment of DDA samples was necessary due to differing gradient lengths but SIL peptides were not required.

Results and Discussion

Searching the Bruderer HeLa DIA data against the DDA library identified 4601 proteins while searching the same data against the Prosit-predicted library resulted in 4221 protein identifications (Figure 4). Searching the yeast data set against the BLIB and Prosit-predicted libraries identified 3433 and 3434 protein identifications respectively (Figure 5).

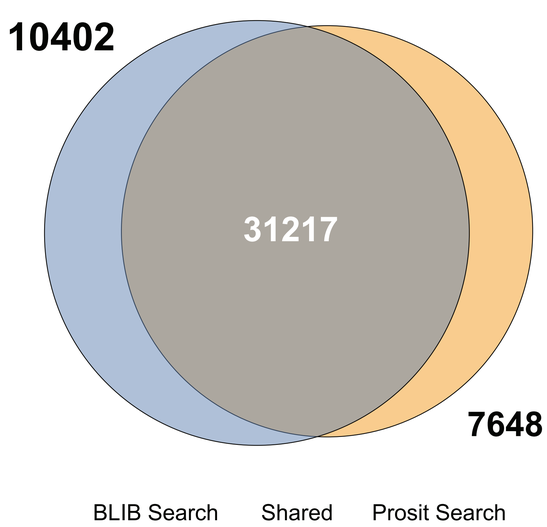

Figure 4: Overlap in peptide detections between library types in HeLa data

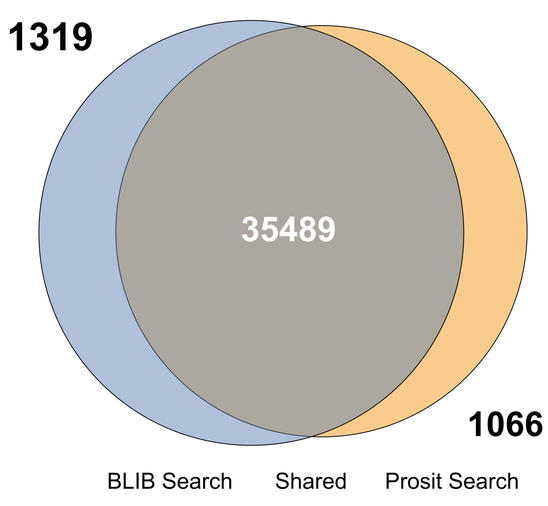

Figure 5: Overlap in peptide detections between library types in yeast data

Researchers now have numerous options when it comes to searching DIA data, each with their own strengths and weaknesses. While the gold standard is to generate a sample-specific DIA library, often this is not feasible due to time or financial reasons. Our work here indicates that existing data can be leveraged in a meaningful way.

Searching the Bruderer HeLa data against a library made of existing fractionated DDA data identified more proteins than searching against a Prosit-predicted library. Additionally, the DDA library came with other advantages. Prosit libraries take time to generate (this can take upwards of a day when using the public server for a large proteome) where existing DDA search results can be loaded into Scaffold, or other library generation tool, and a library can be exported in a few hours. Setting up an in-house server is possible and it often allows for faster library generation, but requires specific hardware, time and technical know-how that many laboratories do not possess.

Discussion (Contd.)

Searching the yeast data against both the Prosit-predicted and DDA libraries identified almost exactly the same number of proteins. In both cases the two approaches produced similar numbers of proteins, within 10% of each other, with the BLIB search identifying more unique peptide sequences.

Scaffold software is ideally suited for this workflow as Scaffold 5 reads both raw data and a wide variety of search engine results and can export a BLIB that is compatible with Scaffold DIA. Both BLIB libraries and Prosit-predicted libraries can be used to search DIA data in Scaffold DIA

Because searching a BLIB or Prosit-predicted library does not require the use of SIL peptides, libraries can be made from a wide variety of existing data files, they did not need to be captured specifically for library generation. This can even include search results from previous experiments and means that libraries can be generated in less time and for less cost as compared to other techniques or programs.

Conclusion

There are multiple ways to search DIA data but each method comes with both benefits and challenges. Creating a sample-specific DIA library will give the best results. However, the cost and time associated with doing so is not always practical. In those cases, laboratories can still get good results by searching a Prosit-predicted library or BLIB library generated from existing DDA data. These libraries can be created in significantly less time and for less cost as additional instrument time is not needed and stable isotope peptides are not required when a library is created using Scaffold.

References

[1] Bruderer, Roland et al. Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Mol Cell Proteomics 16(12): 2296-2309 (2017)*

[2] Searle, Brian et al. Generating high quality libraries for DIA MS with empirically corrected peptide identifications. *Nature Comm. 11: 1548 (2020)*

Contact Information

- Web: www.proteomesoftware.com
- Email: jacob.lippincott@proteomesoftware.com
- Phone: +1 800 944 6027