Triqler for Data Independent Aquisition Data

Patrick Truong Matthew The Lukas Käll

June 18, 2021

Abstract

In this study we show that Triqler, a protein quantification and differential analysis tool based on probabilistical graphical models, has better performance than other protein quantification tools. To show this we compare different processing pipelines using different underlying concept for protein idenfication and quantification...

Introduction

Label-free quantification (LFQ) using Mass spectrometry (MS) based proteomics has been shown to be an effective methods for studying the relative concentration of proteins in complex mixtures. Compared to Data-dependent axquisition (DDA), Data-independent acquisition (DIA) mass spectrometry allows for a broader dynamical range and more reproducible peptide detection ´ zhang2020DIA, Lu2021DIAmeter.

Triqler is a novel software that uses a probabilitical graphical model for protein quantification and differential expression analysis, essentially eliminating the need for filtering, tresholding and imputational procedures required by many conventional methods. Triqler has been shown to distinguish more proteins for DDA data compared to other DDA protein quantification methods. The2018Integrated.

Materials and methods

Data description

The data is a DIA dataset used in a previous benchmark of DIA protein quantification benchmarking study [LFQBenchPaper2016]. It is available from the ProteomeXchange Consortium with the dataset identifier PXD002952. The instrumentation used process the data was TTOF6600 system with 32 fixed windows. In the repository the data we use is referred to as the HYE124 hybrid proteome samples. It consists of tryptic peptides with the following ratios: Sample A composed of 65% w/w, 30% w/w yeast, and 5% w/w E. coli proteins. Sample B was composed of 65% w/w, 15% w/w yeast and 20% w/w E. coli proteins. Further details about mass spectrometric instrumentation and data acquisition is available in Navarro et al. [LFQBenchPape2016].

Data preparation and spectral library generation The .wiff files are converted to .mzML files in centroided format using msconvert (using windows OS msconver version X.X) with the following options: [check options].

Two approaches was used for spectra library generation: DDA acquisition based spectral library generation and Prosit-based spectral library generation using only .fasta file [cite prosit paper].

DDA acquisitions of samples from each specie (human, yeast, E. coli) was provided in triplicates for spectral library generation. Uniprot fasta files with one protein seqeunce per gene was concatenated for each specie (UP000005640, UP000000625 and UP000002311, acquired on 2021-06-16).To control for the effect of different protein inference strategies (protein group, parsimony etc.) a modified .fasta file, without shared peptides, was used for database search. The unfiltered fasta files contained 20 590 human proteins, 6 046 yeast proteins and 4 373 E. coli proteins. After filtering the fasta file contained 20 302 proteins (-288 human proteins), 5 848 yeast proteins (198 yeast proteins) and 4 306 E. Coli proteins (-67 E. Coli proteins). Each sequence with length >7 amino acids mapping only to one protein. The fasta file contained reverse sequences as decoys for target-decoy analysis. MSFragger with parameters: [check parameters] was used for DDA-search, statistical validation was performed by peptide prophet and protein prophet, and EasyPQP with parameters: [check parameters] was used for spectral library building. OpenSwathDecoyGenerator was used with default setting to generate decoys for the resulting spectral libraries.

For Prosit-based spectral library generation, the fasta file was converted to prosit input format using encyclopeDIA converter. Prosit_2020_intensity_cid model was used as intensity prediction model and Prosit_2019_irt was used as iRT predicition model.

OpenSwath Analysis Version (version) of OpenSwath was used. The spectral library generated above is converted to .pqp format using TargetedFileConverter. Data analysis was conducted using OpenSwathWorkflow with parameters (-Scoring:TransitionGroupPicker:background_subtraction original -Scoring:stop_report_after_feature -1, -min_upper_edge_dist 1, -tr_irt hroest_DIA_iRT.TraML, -extra_rt_extraction_window 100, -min_rsq 0.95, -min_coverage 0.6, -Scoring:Scores:use_dia_scores true, -rt_extraction_window 600, -mz_extraction_window 30, -threads 10, -Scoring:DIAScoring:dia_extraction_unit ppm). After data extraction the data .osw output was merged using pyprophet merge option and pyprophet was used for statistical validation. Pyprophet export was used without FDR filtering (–max_global_protein_qvalue 1.0, –max_global_peptide_qvalue 1.0, –max_rs_peakgroup_qvalue 1.0, –max_transition_pep 1.0) to give a complete list of peptide quantifications for further downstream analysis.

DIAUmpire and DIA-NN analysis DIAUmpire signal extraction (SE) was used through Fragpipe GUI (v15.0). Default parameters was set (MS1 PPM: 10, MS2 PPM: 20, Max Missed Scans:1, Mass Defect Filter On, RP max: 25, RF max: 500, Corr Treshold:0, Delta Apex: 0.2, RT Overlap 0.3, Mass Defect Offset 0.1, Isotope Pattern: 0.3, MS1 SN: 1.1, MS2 SN 1.1, Adjust fragment intensity On). MSFragger was used on the resulting .mzML files from DIAUmpire SE with default parameters for Peak Matching (PPM: [-20, 20], Fragment mass tolerance PPM: 20, Calibration and Optimization: Mass Calibration, Parameter optimization, Isotop error: 0/1, Data type: DDA,), protein digestion (Load rules: stricttrypsin, Enzyme name: stricttrypsin, Cut after: KR, Cleavage: ENZYMATIC, Missed cleavages: 2, Clip N-term M: On, Peptide length 7-50, Peptide mass range: 500-5000, Split database: 1) and Modification (Variable modifications: M, [; Fixed modification: "all selected").

EncyclopeDIA and PECAN analysis Prosit was used to contruct spectra libraries from the modified fasta files. The fragmentation model used was "Prosit - Model - Fragmentation" and the iRT model "Prosit - Model - iRT" (available from https://figshare.com/projects/Prosit/35582).

...

Protein quantification

Triqler Triqler was used with –fold_change_eval between 0-2 with 0.04 increments. It computes the two-sided differential probability treshold between the two samples given a fold change evaluation limit.

Top3 The precursors are filtered by q-value ¿ 1% and the average of the three largest peptide intensities are taken for each protein. Protein with only one detected peptides (single hit proteins) and proteins detected only in two injections are discarded.

Msstat

Msqrobsum

Results

OpenSwath Analysis

DIAUmpire and DIA-NN analysis

EncyclopeDIA and PECAN analysis

Discussion

Acknowledgements

Funding

This work has been supported by a grant from the Swedish Foundation for Strategic Research (BD15-0043).

Supporting information

References