Performance of LPGF Protein Validation in diverse sample types and mass spectrometry workflows

Pedro Navarro[1], Waqas Nasir[1], Kai Fritzemeier[1], Gorka Prieto[2], Víctor M. Guerrero-Sánchez[3], Jesús Vázquez[3], Christoph Henrich[1]

1. Thermo Fisher Scientific (Bremen) GmbH, Bremen, Germany 28199
2. Dept. of Communications Engineering, Univ. of the Basque Country (UPV-EHU), Bilbao, Spain 48013
3. Centro Nacional de Investigaciones Cardiovasculares Carlos III, Madrid, Spain 28029

Abstract

Purpose: Evaluation of the performance of a new protein validation algorithm (LPGF) recently added to Proteome Discoverer.

Methods: The previous protein validation nodes from Proteome Discoverer were compared with the newly introduced LPGF nodes using various techniques. We conducted an evaluation of the False Discovery Rate (FDR) using entrapment databases to assess the precision of the FDR. Standard databases were utilized to evaluate the performance of identification. Additionally, we assessed the value of newly identified proteins by examining the precision and consistency of quantification in multi-proteome mixes and dilution series datasets.

Results: The new LPGF algorithm identifies from 3% to 9% more proteins, depending on the analyzed dataset. The proteins discovered uniquely by LPGF show consistent quantification.

Introduction

Validation of protein identifications in large-scale mass spectrometry proteomics datasets is a long-pursued challenge. In the target-decoy validation strategy, the accumulation of decoy candidates complicates the separation of target and decoy score distributions. Protein FDR is also considerably increased because FDR-validated peptides contribute to multiple proteins, making it critical the use of appropriate protein scoring algorithms. LPGF Protein Validation was firstly introduced in 2020 (Ref. 1) as a novel protein scoring method based on a highly accurate estimation of decoy identification probabilities. The model also includes a refinement of the "picked" FDR estimation method.

Materials and methods

Sample Selection and Data Analysis

MS-prepared peptidome datasets comprising several sample types: human cell lines (HeLa), human plasma, and human phospho-enriched peptidome, and several mass spectrometry workflows including data dependent and data independent acquisition performed in Thermo Scientific™ QExactive™, QExactive™+, QExactive™ HFX™, Exploris™ 480, and Orbitrap™ ASTRAL™ (details in table 1) were processed in Thermo Scientific™ Proteome Discoverer™ 3.1 (PD) with SEQUEST HT (DDA data) and CHIMERYS 2.7.9 (DDA & DIA data) with canonical fasta databases and standard modifications (fixed Cys-CAM and variable Met-Oxidation). The Protein validation was performed by using the standard PD 3.1 validation nodes (Protein Scorer and Protein Validator), and by two new Proteome Discoverer nodes (Protein Scorer LPGF and Protein Validator LPGF) developed for this study.

Table 1. List of tested datasets

Instrument	Acq. Method	Organism	Sample amount	# raw files	Gradient [min]
Orbitrap ASTRAL	DIA	HeLa	200 ng	3	24
Orbitrap QE+	DDA	HeLa	1 µg	3	30, 60, 120
Orbitrap QE HFX	DDA	HeLa	1 µg	2	75
Orbitrap ASTRAL	DIA	Human Plasma	200 ng	3	15
Orbitrap ASTRAL	DIA	HEK239T Phopho	250 ng	3	7, 15, 30
Orbitrap ASTRAL	DIA	Human, Yeast, E.coli	1 µg	6	30
Orbitrap Exploris 480	DIA	Human, Yeast, E.coli	500 ng	6	45
Orbitrap Exploris 480	DIA	Human	10 ng to 0.06 ng	24	15
Orbitrap ASTRAL	DIA	Human	20 ng to 250 pg	36	15

Results

Identification Performance

We first compared the number of protein identifications achieved by both former PD nodes and the new LPGF nodes in datasets. Entrapment databases were used to check that reported FDR matches actual FDR values (right side of Fig. 1).

Figure 1. Dataset of 200 ng HeLa analyzed in Orbitrap ASTRAL (3 technical replicates, DIA acquisition, 24 min. gradient). Target proteins discovered vs FDR (left plot) show 5% more protein identifications for LPGF (red line) compared to former PD nodes (blue) at 1% FDR. FDR entrapment vs reported FDR (right plot) shows very similar FDR control for both methods. The FDR entrapment plots show very similar look for all analyzed datasets.

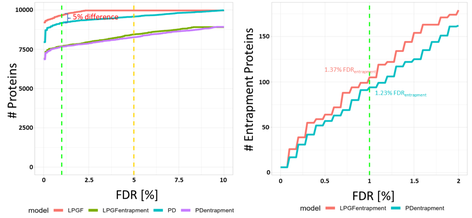
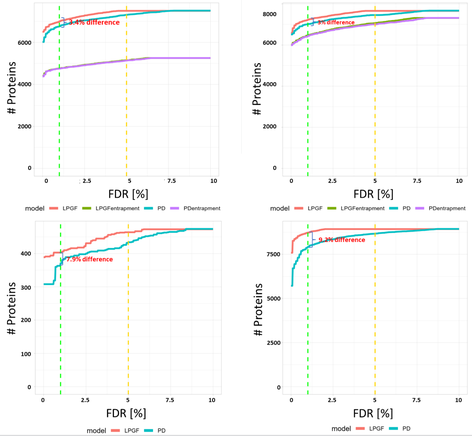


Figure 2. 1 µg HeLa analyzed in DDA acquisition 24 min. gradient in an Orbitrap QE+ (up left plot). 1 µg HeLa in QE DDA, 3 raw files with 30-, 60-, and 90-min gradients respectively (up right). 1 µg HeLa in QE HFX DDA (2 replicates 75 min gradient) (down left). 200 ng Human plasma (down left) and 250 ng HEK239T phospho-enrichment (down right).



Quantification Precision

We benchmarked quantification precision on triple proteome mix samples (HeLa, Yeast, and E. coli). Scatter plots of ratios vs protein quantities show that uniquely identified proteins by LPGF tend to be at the lower amount area (as expected). The quantification precision matches the expectations for low amount proteins (Figs. 3 & 4).

Figure 3. 1 µg Triple proteome mix measured in Orbitrap ASTRAL (30 min gradient). Top plot shows all proteins quantified in both conditions; bottom plot shows proteins uniquely quantified.

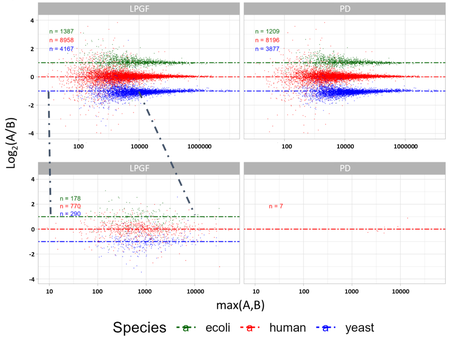
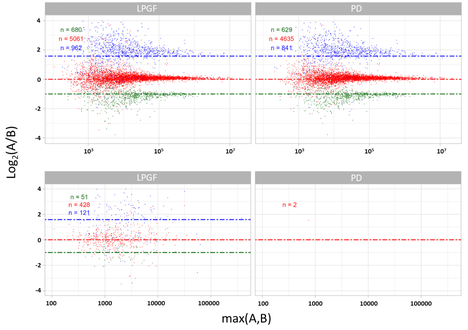


Figure 4. 500 ng Triple proteome mix measured in Orbitrap Exploris 480 (45 min gradient). Top plot shows all proteins quantified in both conditions; bottom plot shows proteins uniquely quantified.



Dilution Series

A dilution series permits to evaluate whether identified proteins show a consistent quantification through the different dilutions. This consistency can be measured by the linear response (see Fig. 5) and by the number of gaps (missing identifications) in the middle of the dilution series (see Figs. 6 & 7).

Figure 5. 10 ng HeLa sample was diluted up to 0.06 ng and analyzed in Orbitrap Exploris 480. For each protein, its maximum detected amount has been normalized to match the maximum expected protein abundance (10 ng). Plots are faceted by the maximum abundance quantified in low, medium, and high abundant proteins. Each line represents response values (abundance) ves expected abundance (dilution). The linear response is represented by the line color (in red proteins with a linear response of $R^2 \geq 0.99$, in blue $0.96 \leq R^2 \leq 0.99$, and in grey $R^2 \leq 0.96$. The right-hand side plot shows uniquely detected proteins.

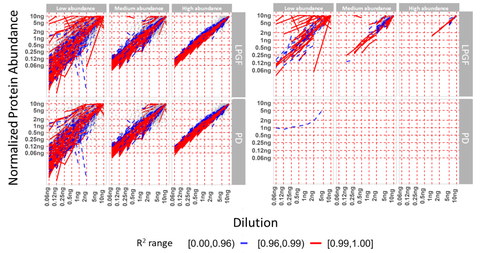
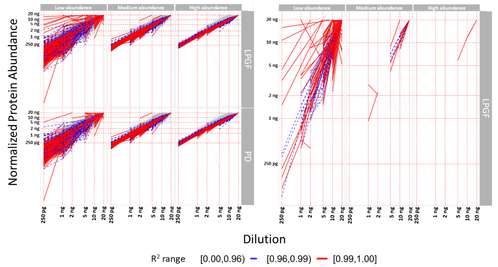


Figure 6. 20 ng HeLa sample was diluted up to 250 pg and analyzed in Orbitrap ASTRAL. For each protein, its maximum detected amount has been normalized to match the maximum expected protein abundance (10 ng). Plots are faceted by the maximum abundance quantified in low, medium, and high abundant proteins. Each line represents response values (abundance) vs expected abundance (dilution). The linear response is represented by the line color (in red proteins with a linear response of $R^2 \geq 0.99$, in blue $0.96 \leq R^2 \leq 0.99$, and in grey $R^2 \leq 0.96$. The right-hand side plot shows uniquely detected proteins.

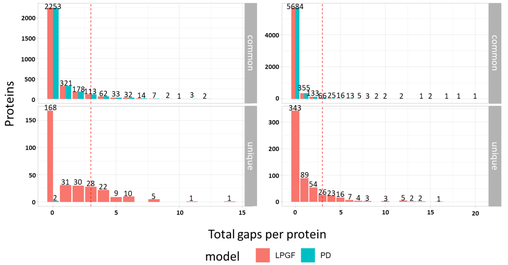


Dilution Series

We have devised a penalty system to assess the presence of gaps within a dilution series of protein concentrations. This series comprises samples at n distinct concentrations, each measured in triplicate.

The evaluation commences from the most diluted sample and progresses towards the one with the highest concentration. The detection of a protein at a given concentration sets the expectation for its detection in an equal or greater number of replicates at higher concentrations. Any deviation from this expectation, manifested as a missing replicate, is classified as a gap and incurs a penalty in our system.

Figure 7. Human dilution series acquired in DIA mode, in Orbitrap Exploris 480 (10 ng to 0.06 ng) (left-hand side), and in Orbitrap ASTRAL (20 ng to 250 pg) (right-hand side). The vertical dashed line represents the maximum acceptable number of gaps. Most LPGF uniquely identified proteins lay below this minimum threshold.



Conclusions

- The False Discovery Rate (FDR) is effectively managed by the LPGF algorithm, as demonstrated by entrapment experiments.
- Proteins identified exclusively by the LPGF algorithm exhibit consistent quantification values, indicating that they constitute a valuable enhancement to the process.

References

1. Protein Probability Model for High-Throughput Protein Identification by Mass Spectrometry-Based Proteomics G. Prieto and J. Vázquez, Journal of Proteome Research 2020 19 (3), 1285-1297

Trademarks/licensing