Implementation of MSFragger and Philosopher (PeptideProphet) as Proteome Discoverer Nodes

Hui-Yin Chang, Andy T. Kong, Felipe V. Leprevost, Fengchao Yu, Guo Ci Teo, Dmitry M. Avtonomov, Venkatesha Basrur, Alexey I. Nesvizhskii

University of Michigan, Ann Arbor, Michigan, United States

Introduction

Goal: Implementation of MSFragger and Philosopher (PeptideProphet) as Proteome Discoverer 2.2 and 2.3 (PD v2.2 and v2.3) nodes.

Input: Files in mzXML/mzML (without zlib compression)/raw file format

Process: MS/MS spectra are first filtered and imported to an MSF file in the SpectrumSelector node. Next, the MSFragger node automatically executes database search on those input files and imports PSMs to the same MSF file. If users select PeptideProphet as the PSM validator, a process will be performed to update the confidence scores of validated PSMs. A typical PD consensus workflow can be applied for downstream analyses to generate final reports.

Output: The tables of identified PSMs, peptides, and proteins along with the reliable statistics and validate quantification results.

Processing Workflow

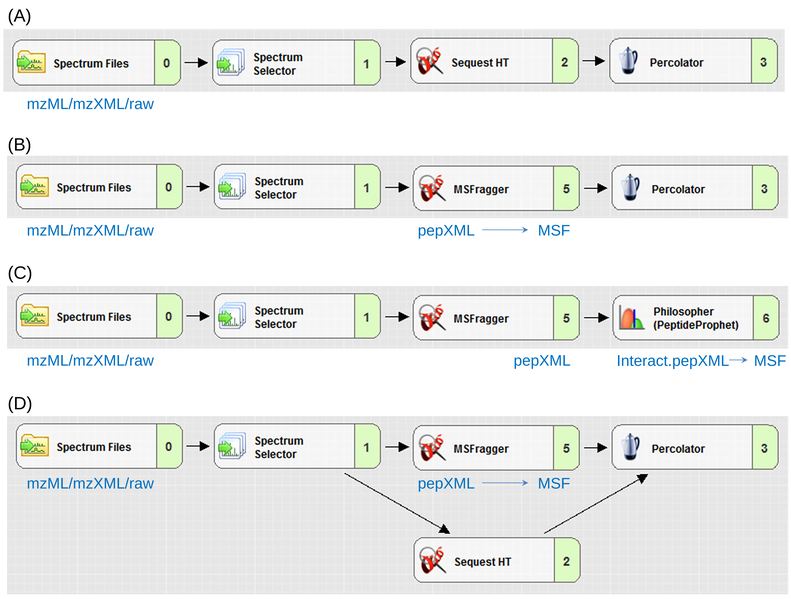

Figure 1. Four processing workflow examples in PD v2.3. (A) A conventional processing workflow using SequestHT. With our MSFragger-PD node, users can either use Percolator as the PSM validator to evaluate MSFragger's results **(B)**, or they can use PeptideProphet as the PSM validator **(C)**, which is suitable for validating open search results. Users can also combine multiple search engines (e.g., SequestHT or MSAmanda) **(D)** for peptide identification. However, only Percolator which is able to directly access MSF files can be used as the PSM validator when combining multiple search engines.

Results

Two proteome data sets were used for performance evaluation: a label-free HEK293 data set (PXD001468), and a labeling data set (PXD008952) where seven cancer cell lines were labeled with tandem mass tag (TMT). All raw files were converted to mzML format using ProteoWizard and imported to PD v2.2 for protein identification and quantitation.

Table 1. The number of identified PSMs, peptides, and proteins (filtered by 1% FDR) using two LC-MS/MS datasets with four different search strategies. We also used FragPipe v9.1 to run MSFragger and Philosopher for comparison. *When combining multiple search engines, a PSM may be repeatedly reported by more than one search engine, and therefore the number of identified PSMs using multiple search engines was higher than using a single search engine.

(A) Label-free HEK293 data set (PXD001468)

Number	**SequestHT/ Percolator**	MSFragger/ Percolator	**MSFragger/ PeptideProphet**	MSFragger/SequestHT/ Percolator	FragPipe v9.1
PSMs	**450261**	532956	**523537**	984015*	550581
Peptides	**123603**	128494	**128892**	134533	125493
Proteins	**8966**	9383	**9223**	9241	9313

(B) TMT NCI-7 data set (PXD001468)

Number	**SequestHT/ Percolator**	MSFragger/ Percolator	**MSFragger/ PeptideProphet**	MSFragger/SequestHT/ Percolator	FragPipe v9.1
PSMs	**326627**	378462	**366653**	696321*	384658
Peptides	**190319**	180794	**175922**	194437	165734
Proteins	**10670**	10991	**10933**	10965	10833

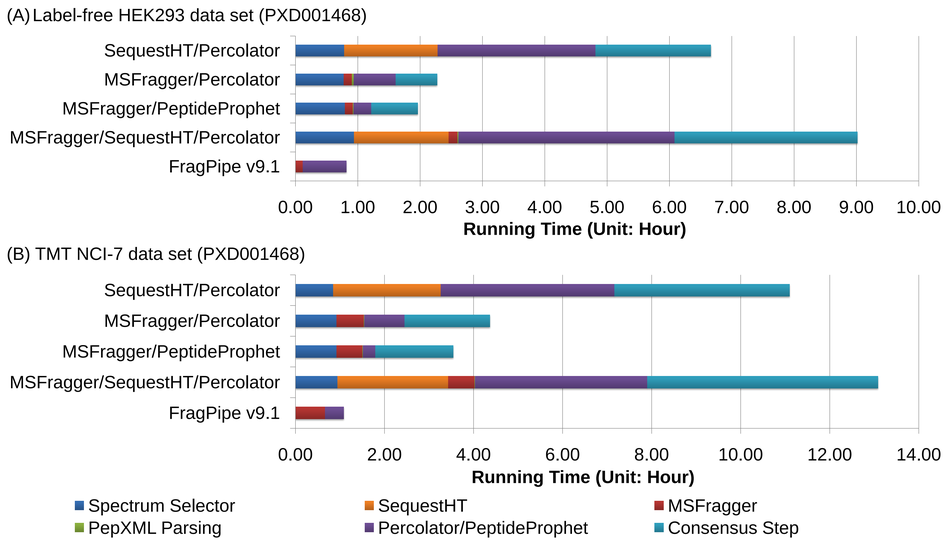

Figure 2. The processing speed of label-free HEK293 data set (all 24 fractions) (A) and TMT NCI-7 data set (all 24 fractions) (B). Note that the processing speed of our MSFragger-PD node is much faster than that of the conventional PD processing and consensus workflows.

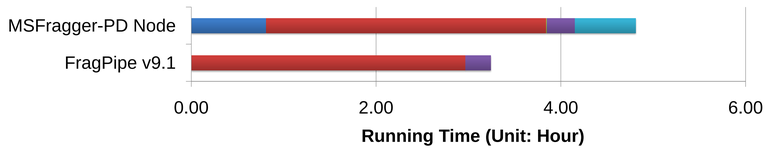

Figure 3. The comparison of processing time between MSFragger-PD Node and FragPipe v9.1 using label-free HEK293 data set (all 24 fractions). Note that the running time of MSFragger and PeptideProphet is similar inside and outside of PD v2.2

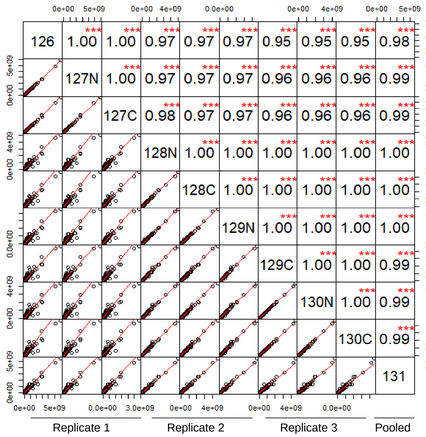

Figure 4. The protein abundance correlation of TMT labeling data set using our MSFragger-PD node. Channel 126, 127N, and 127C are from replicate 1, channel 128N, 128C, and 129N are from replicate 2, and channel 129C, 130N, and 130C are from replicate 3, and channel 131 is from the pooled sample. High correlations (above 0.9 in average) were also observed in the search results using SequestHT/Percolator, MSFragger /Percolator, and MSFragger/Sequest-HT/Percolator.

Summary:

Our MSFragger-PD node:

- Identifies more PSMs, peptides, and proteins than conventional PD workflow using SequestHT
- Has fast processing speed
- Enables open search in PD
- Provides accurate quantitation results
- Fully cooperates with downstream nodes/workflows in PD v2.2 and v2.3.

Our Posters Open-Search Comparison (MP402), MSFragger (MP 405), FragPipe (MP 416), MSFragger-PD Nodes (MP437), PTM-Shepherd (ThP693), Philosopher (WP 396)

Acknowledgement We would like to thank Dr. David M Horn and Carmen Paschke for helping us developing and testing our MSFragger-PD node, and Dr. Karthik S Kamath, Dr. Amanda Buchberger, Dr. Rebekah L. Gundry, Dr. Joao A Paulo, Dr. Karl Mechtler, and Dr. Otto Hudecz for helping us testing the node.

This work was supported by NIH Grants R01-GM-094231 and U24-CA210967

Learn more : *https://github.com/Nesvilab/PD-Nodes*