Implementing Comet search engine into Proteome Discoverer to improve TMT Real-Time Search data processing

[1] Yang Liu, [2]Frank Berg, [1] William D. Barshop, [1]Jesse D. Canterbury, [1]David Horn, [1]David Bergen, [1]Romain Huguet, [1]Rosa Viner
[1]Thermo Fisher Scientific, San Jose, CA, USA. [2]Thermo Fisher Scientific, Bremen, Germany

Overview

Purpose: To compare different search engines and PSM validation approaches in Thermo Scientific™ Proteome Discoverer™ software, in order to achieve the best alignment between real-time database search and post-acquisition data analysis.

Methods: Thermo Scientific™ Pierce™ TMT11plex Yeast Digest Standard and TMT18plex Yeast Digest Standard (prototype) were analyzed by Thermo Scientific™ Orbitrap Eclipse™ Tribrid™ mass spectrometer. Real-time database search was performed by Comet algorithm and post-acquisition data analysis was carried out using PD 3.0.

Results: Comet algorithm in Proteome Discoverer 3.0 provided better alignment to real-time search data acquisition compared to the Sequest® HT search engine. Fix value validator is the optimal choice in PSM validation without FDR calculation. Percolator node is recommended if FDR threshold is needed. The combination of Sequest HT and Comet in Proteome Discoverer software improved the identifications and quantification IDs.

INTRODUCTION

Real-time search (RTS) using Comet on the Thermo Scientific™ Orbitrap Eclipse™ Tribrid™ has enabled selective triggering of SPS MS3 scans upon confident identifications from MS2 spectra. This method largely improved the identification numbers and quantitation accuracy. The most common post-acquisition data analysis search engine for such data is Sequest HT, the primary search engine in the PD software. Although Sequest HT and Comet share a similar heritage, discrepancies may still exist in spectral processing and interpretation. Here we introduce the implementation of Comet in Proteome Discoverer 3.0 to provide the best alignment between online and post-acquisition data analysis. Moreover, we compare multiple FDR validation approaches in Proteome Discoverer and provide suggestions for the optimal choice in different conditions.

MATERIALS AND METHODS

Sample Preparation

TMT11plex Yeast Digest Standard was reconstituted to a final concentration of 250ng/ul in 0.1% TFA/5% acetonitrile in LC/MS-grade water. TMTpro 18plex Yeast Digest Standard (prototype) was reconstituted to a final concentration of 250ng/ul in 0.1% TFA/5% acetonitrile in LC/MS-grade water.

Test Method(s)

The samples were analyzed by an Orbitrap Eclipse Tribrid mass spectrometer (ICSW 3.5) coupled to the Thermo Scientific™ EASY-nLC™ 1200 chromatography with a 50 cm Thermo Scientific™ EASY-Spray™ Column. Yeast digest peptides were separated at 50min and 120min LC gradient before the injection to mass spectrometer. MS2 spectra were online searched against a yeast proteome database during acquisition using the Comet search algorithm (2019.01 rev.1). The real-time search scoring thresholds were: Xcorr 1.4, dCn 0.1 and mass tolerance 10ppm.

Data Analysis

The data were analyzed with the beta version of Proteome Discoverer 3.0, using both Comet and Sequest HT search algorithms. The search parameters in the software were matched to the RTS Comet settings. Carbamidomethylation was considered static modification and oxidation (M) was dynamic modifications. TMT and TMTpro tags were set up as static modifications at Lysine and N-terminal. Multiple PSM validation nodes, such as Percolator, fixed value PSM validator and target decoy PSM validator were tested to find the best alignment with the RTS Comet search result.

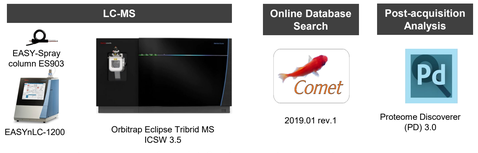


Figure1. Design and workflow of RTS data acquisition and data analysis

RESULTS

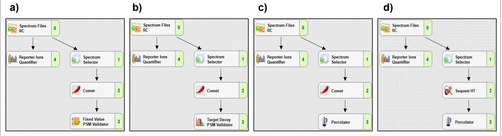
The same RTS raw file was analyzed by different Proteome Discoverer workflows using multiple PSM validators (Figure 2).

Fix Value PSM validator: Perform validation of PSMs based on score threshold defined for the search node;

Target Decoy PSM validator: Perform validation of PSMs based on score threshold derived from target/decoy result to meet a specific target false discovery rate.

Percolator: Calculate posterior error and probabilities and q-values for the identified PSMs using Percolator

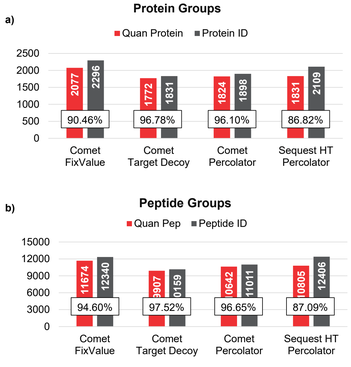
Figure 2. Proteome Discoverer workflows using Comet algorithm with a) Fixed Value PSM validator, b) Target Decoy PSM validator, c) Percolator and d) Sequest HT search engine with Percolator



With different PSM validation strategies, Proteome Discoverer analysis resulted in different number of identified and quantified protein groups and peptide groups. In Fix Value PSM Validator, delta Cn was set up as 0.1, which is same as that was used in online database search. In Target Decoy PSM Validator and Percolator, peptide target FDR was setup as 1%. The comparison of results are shown in Figure 3. Comet Fix Value PSM Validator resulted in the highest protein groups and peptide groups, both identification and quantified IDs. Comet Target Decoy PSM validator workflow resulted in the highest quantification percentage, However, it generated the fewest identified and quantified protein groups and peptide groups.

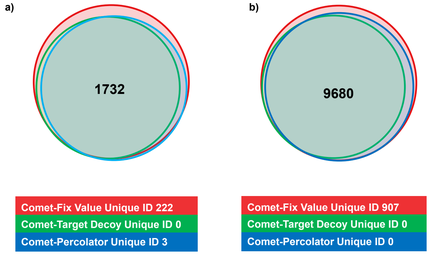
500ng TMT 11plex Yeast Digest Standard (50min LC gradient)

Figure 3. Protein groups and peptide groups identified and quantified from TMT11plex Yeast digest standard using different PD analysis workflows



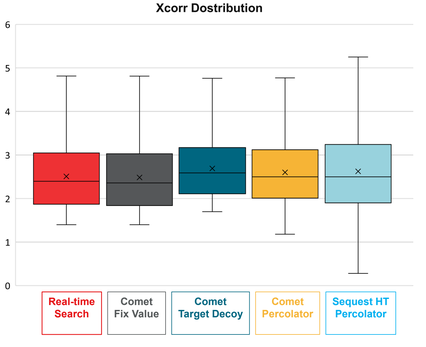
In order to confirm the identification coverage, we imported the quantified protein groups and quantified peptide groups from different PSM Validators for a Venn Diagram comparison (Figure 4). Since Fix Value PSM Validator did not have any FDR filter, it resulted in the broadest coverage.

Figure 4. Venn Diagrams to compare Quantified a) protein groups and b) peptide groups



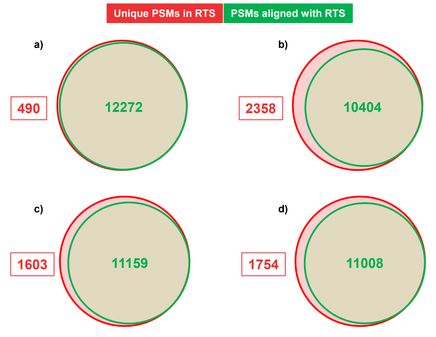
In addition to the number of identifications and quantified IDs, we would like to get the best alignment between online real-time search and post-acquisition data analysis. The Xcorr distributions of PSM identified in real-time database search and each Proteome Discoverer analysis workflow are shown in Figure 5. We can tell from the boxplot that Comet Fix Value PSM Validator generated the best alignment with online search in Xcorr distribution.

Figure 5. Xcorr Distribution of PSMs identified in different Proteome Discovere workflows.



In SPS-MS3 RTS method, quan spectra (MS3) was only triggered upon a confident PSM identification (MS2). Therefore, ideally, quantified PSMs in Proteome Discoverer analysis result should align with the MS2 scans which have passed the user-defined RTS threshold. Here we plotted the venn diagram (Figure 6) to illustrate the alignment of online and post-acquisition data analysis results. The best alignment was achieved by Comet Fix Value PSM Validator.

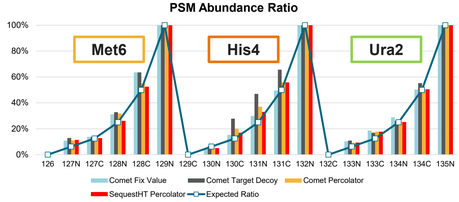
Figure 6. Venn Diagrams of PSM quantified by Real-time Search and a) Comet Fix Value b) Comet-Target Decoy c) Comet-Percolator d) Sequest HT-percolator.



500ng TMT 18plex Yeast Digest Standard Prototype (120min LC gradient)

Recently we have developed a new standard sample, TMT 18plex Yeast Digest Standard (prototype), in which the knock-out channels (Met6, His4 and Ura2) were mixed in different ratios with parental (ENO2). Expected ratio in each TMT channel is shown as the solid line in Figure 7 This new TMT standard would be a great example in evaluating the performance in both ID discovery and quantitation accuracy. The same RTS raw file was analyzed by Comet algorithm in PD combined with three different PSM validator nodes, and Sequest HT Percolator as well. Channels 129N, 132N and 135N were used as control. The most distorted ratio was generated from Comet Target Decoy workflow. In most of the quan channels, Comet Fix Value and Comet Percolator were able to achieve similar accuracy. Sequest HT Percolator performed the best in alignment to the expected ratios.

Figure 7. Quantification Accuracy of TMT18plex Yeast Digest Standard Prototype using different PD analysis search engines and PSM validation methods.



In order to identify and quantify more IDs while maintain the quantification accuracy and alignment with online database search, we combined both Sequest HT and Comet in a parallel PD analysis workflow (shown in Figure 8). By taking advantages of the complementary coverages from two search engines, both identified and quantified protein groups have been improved 5-10% compared to single search algorithm. (shown in Figure 9)

FIGURE 8. Proteome Discoverer workflow combining Comet and SequestHT algorithm

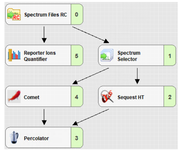
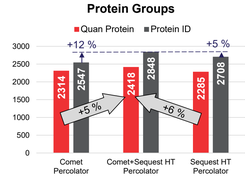


Figure 9. Protein groups identified and quantified TMT18plex Yeast digest standard using different PD analysis workflow



CONCLUSIONS

- Sequest HT and Comet are two search engines that share a similar heritage, however, discrepancies may still exist in spectral processing and interpretation, which leads to the misalignment of real-time search data acquisition and post-acquisition data analysis.
- For the post acquisition data analysis of RTS data, Proteome Discoverer workflow using Comet algorithm and Fix Value PSM Validator provided the best alignment with online database search result.
- If an FDR threshold is desired in data analysis, Proteome Discoverer Comet coupled with Percolator is recommended.
- The Proteome Discoverer workflow using Sequest HT Percolator resulted in the best quantification accuracy in analyzing TMT18plex Yeast Digest Standard Prototype.

TRADEMARKS/LICENSING

PO66113 EN0921S