Integration of mProphet chromatogram peak identification probability model into Skyline

Brendan MacLean¹, Don M. Marsh², Hannes Röst³, Lucia Espona Pernas², Olga Schubert⁴, George Rosenberger⁵, Ruedi Aebersold⁶, Michael J. MacCoss⁷

¹University of Washington, Department of Genome Sciences, ²Institute of Molecular Systems Biology, ETH Zurich

The Skyline Targeted Proteome Environment has distinguished itself as a reliable and useful tool in chromatography-based quantitative proteomics. From its initial focus on selected reaction monitoring (SRM) to its current support for high-sensitivity methods including MS/MS based targeted MS/MS and data independent acquisition (DIA), specifically for the analysis of the characteristic features of a proteome, hundreds of publications and thousands of documents have been based on Skyline under various experimental conditions. Until now, however, peak identification within Skyline has relied on a limited set of features without the ability to derive a statistical confidence metric like a false discovery rate (FDR). To correct this, we have now integrated the mProphet scoring method into Skyline, allowing for improved identification and derivation of peak q-values within Skyline.

Methods:

The integrated mProphet scoring method requires a list of transitions derived from chromatograms. These lists of transitions are often generated from references or decoy databases and contain a variety of information from the original references, such as run information, retention time, mass-to-charge ratio, and precursor ion. By integrating these lists into Skyline and applying the mProphet scoring method, a q-value is derived for each transition, providing a measure of how confident the transition is identified. This method has been shown to improve the separation of true and false transitions, resulting in more accurate and reliable peak identification in Skyline.

Results:

Figure 1: Hypothetical chromatogram and resulting simplified models as well as the existing Skyline peak scoring methods.

In order to achieve this, we have implemented the following:

1. Deriving a set of transitions from two different databases:
   - Monocore peptide and transition generation
   - Decay peptide and transition import

2. Classification and regression training using a Support Vector Machine (SVM) classifier:
   - Summary - requiring only summary peak attributes
   - Detailed - requiring chromatographic information

3. Calculating peak q-values using a combined scoring function during import:
   - mProphet_q
   - Malmstrom_q
   - Intensity correlation (dot-product)

4. Calculating several new scores:
   - Malmstrom_q
   - Intensity correlation

5. Calculating a new combined score algorithm:
   - Skyline q-value calculation algorithm

Integrating the mProphet scoring method into Skyline provides several advantages:

- Improved separation of true and false transitions
- Increased accuracy and reliability of peak identification
- Reduced implementation effort

Introduction:

Skyline is a powerful tool for the analysis of mass spectrometry data, allowing for a variety of analyses such as chromatogram peak identification, feature discovery, and target validation. However, until now, the software did not include a statistical confidence metric like a false discovery rate (FDR). To address this gap, we have integrated the mProphet scoring method into Skyline, enabling a more comprehensive approach to peak identification within the software.

Figure 2: Foremost, reporting peak groups and adding them to PEP files for use with Skyline (right).

Conclusions:

- Good progress has been made in integrating the mProphet scoring model into Skyline in a way that makes important information for assessing model accuracy immediately available.
- Much work remains to be done to understand how much scoring separation will enhance peak picking in Skyline and whether adding further cores is required for improving its performance.
- Overall, the implementation of mProphet into Skyline is a significant step forward for chromatogram peak identification.

References: