# ProMex: a New Feature Extraction Algorithm for Top-down Mass Spectrometry based on Summed Spectra

Jungkap Park, Paul D. Piehowski, Christopher S. Wilkins, Anil K. Shukla, Yufeng Shen, Samuel H. Payne, Richard D. Smith and Sangtae Kim

Biological Sciences Division, Pacific Northwest National Laboratory, WA

Pacific Northwest NATIONAL LABORATORY

*Proudly Operated by* Battelle *Since 1965*

## Overview

- ProMex: **Pr**otein **M**ass **Ex**tractor
- A new feature finding tool for extracting masses, charge states, elution times, and abundances of intact proteins in top-down mass spectra
- It clusters isotopic envelope peaks distributed across multiple charge states, scans and natural isotopes
- Extracted features are used for protein identification, characterization of post-translational modifications and quantification of proteoforms.

## Introduction

- Top-down mass spectra are complex and noisy
  - Low signal-to-noise ratio
  - Complex mixtures of hundreds or thousands proteoforms

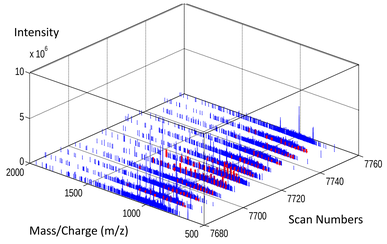



**Figure 1.** An example of top-down LC-MS spectra. Marked red are isotopic envelope peaks from a 28K Da intact protein. They are eluted over a period of time with multiple charge states.

## Methods

### Clustering Isotopic Envelopes

- ProMex clusters isotopic envelopes across multiple charge states and scans

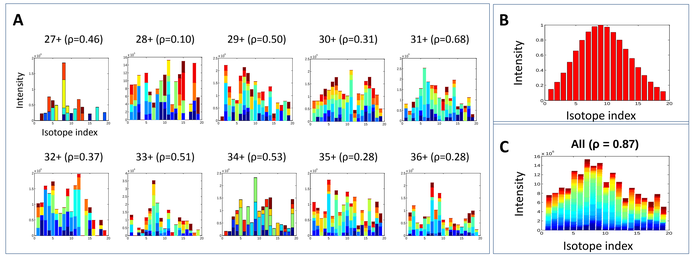



**Figure 2.** (A) Summed isotopic envelopes of the 28K protein in Figure 1 across multiple scans (12 spectra) at each charge state between 27+ and 36+. Numbers in parenthesis are Pearson's correlations to theoretical isotopic envelope, (B) Theoretical isotopic envelope, and (C) Sum of all isotope envelopes in (A).

- Summed envelopes becomes more similar to their theoretical envelopes

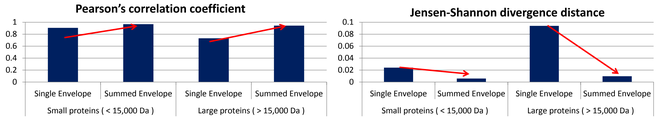



**Figure 3.** Summing envelopes increases similarity to theoretical isotopic envelopes, in particular for large proteins (e.g. >15K Da). Pearson's correlation coefficient and Jensen Shannon divergence distance are used for measuring the similarity between experimental and theoretical envelopes.

### Scoring

- Extracted features are evaluated and filtered out using various scoring metrics

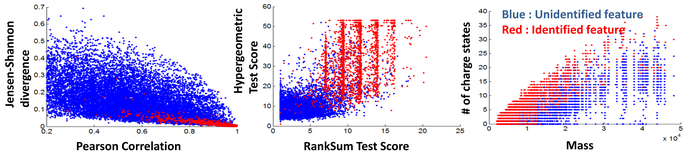



**Figure 4.** Scoring metrics such as Pearson's correlation, Jensen-Shannon divergence, Hypergeometric test score, Ranksum test score and differences in m/z values can separate good and bad features.

## Results

### Protein Identification

- Features extracted by ProMex are used for proteoform identification
- Increase in number of identifications

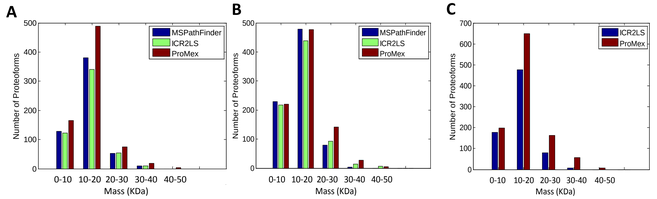



**Figure 5.** The number of unique proteoforms identified at 1% false discovery rate for three datasets from Shewanella oneidensis samples. (A) Short gradient C2 RPLC; Gradient Length = 180 min, (B) Long gradient C2 RPLC; Gradient Length = 850 min, and (C) Long gradient C2 RPLC; Gradient Length = 900 min

- ProMex extracts features of large intact proteins correctly

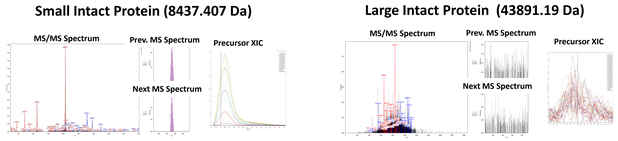



**Figure 6.** Examples of identified proteoforms: MS/MS spectra, adjacent MS spectra, and Precursor XICs of natural isotopes. Matched b ion, y ion and isotopic envelope peaks are marked blue, red, and purple, respectively.

### Label-free Quantification

- Discriminatory features are identified by selecting the features whose abundances are significantly different among multiple groups of samples

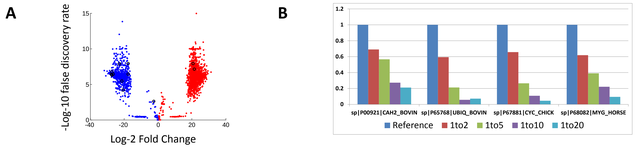
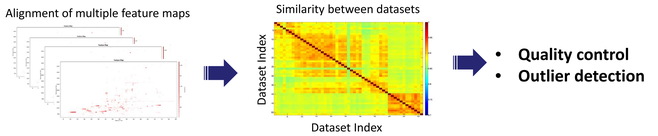



**Figure 7.** A known amount of a set of 4 intact protein standards were spiked into a ovarian tumor lysate background. A statistical analysis on extracted features identifies the spike-in proteins with their correct abundance levels. (A) volcano plot to compare spike-in samples. All spike-in proteins are detected above a 1% false positive threshold (B) The abundance of features corresponding to the spike-in proteins matches well to the spike-in levels

*(Note: Figure 7 will be updated with new data)*

## Measuring Reproducibility

- The similarity of aligned feature maps represents the overall difference between samples




## Discovering Unknown Modifications

- Infer unknown modifications from frequent mass differences between co-eluted features

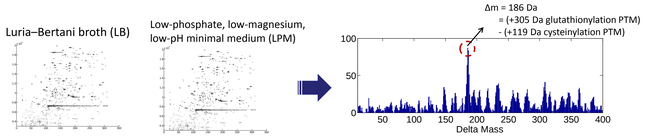



**Figure 8.** Protein S-thiolation switch in Salmonella Typhimurim in response to infection like (LPM) conditions versus basal (LB) conditions. The most frequent mass difference between aligned features is 186 Da, which corresponds to S-glutathionylation-to-Scysteinylation switch. Datasets used in Ansong et al., 2013

## Conclusions

- ProMex is a new feature finding tool which clusters ion peaks scattered over multiple charge states and spectra to better define and allow identification of isotopic envelopes.
- Compared to existing deconvolution algorithms, it increased the number of proteoform identifications by more than 10%.
- In particular, ProMex significantly increased proteoform identifications with masses over 25,000 Da, which were mostly missed by other tools.
- ProMex can be used for label-free quantification of intact proteins.

## Acknowledgements

This work was funded by CPTAC (U24CA160019). Samples were analyzed using capabilities developed under the support of NIH National Institute of General Medical Sciences (8 P41 GM103493-10). This project was supported by the U.S. Department of Energy (DOE) Office of Biological and Environmental Research (OBER) Pan-omics program at Pacific Northwest National Laboratory (PNNL) and performed in the Environmental Molecular Sciences Laboratory, a DOE OBER national scientific user facility on the PNNL campus. PNNL is a multiprogram national laboratory operated by Battelle for the DOE under contract DE-AC05-76RL01830.

**Career Opportunities:** For potential openings in Integrative Omics Group at PNNL please visit http://omics.pnl.gov/careers

**CONTACT: Jungkap Park, Ph.D.**
Biological Sciences Division
Pacific Northwest National Laboratory
**E-mail: jungkap.park@pnnl.gov**

**www.omics.pnl.gov**