A.11
An AI-driven leap forward in peptide identification through deconvolution of chimeric spectra

Martin Frejno1, Daniel P. Zolg1, Tobias Schmidt1, Siegfried Gessulat1, Michael Graber1, Florian Seefried1, Magnus Rathke-Kuhnert1, Samia Ben Fredj1, Patroklos Samaras1, Kai Fritzemeier2, Frank Berg2, Waqas Nasir2, David Horn3, Bernard Delanghe2, Christoph Henrich2, Bernhard Kuster4, Mathias Wilhelm4

1MSAID GmbH, Garching b.München, Germany, 2Thermo Fisher Scientific (Bremen) GmbH, Bremen, Germany, 3Thermo Fisher Scientific, San Jose, CA, 4Technical University of Munich, Freising, Germany

Objective: Chimeric spectra are estimated to constitute $\geq$40% of DDA data (Dorfer et al., 2018), violating the assumption that one spectrum represents one peptide. Here, we describe a new intelligent search algorithm (CHIMERYS) that rethinks the analysis of tandem mass spectra from the ground up. It routinely doubles the number of peptide identifications and reaches identification rates of $\geq$80%.

Methods: Our new algorithm uses accurate predictions of peptide fragment ion intensities and retention times provided by a deep learning framework. All candidates in the isolation window of a given tandem mass spectrum are considered simultaneously and compete for measured fragment ion intensity in one concerted step. The algorithm aims to explain as much measured intensity with as few candidate peptides as possible, resulting in the deconvolution of chimeric spectra. FDR-control is performed using Percolator (The et al., 2016). Searches can be triggered from laptops and conventional workstations (via a node in Proteome Discoverer 3.0 software) and are parallelized in the cloud.

Results: Analyzing a HeLa trypic digest (1 hour gradient) with our new algorithm identified 114k PSMs, 61k unique peptides and 7,300 unique protein groups at 1% FDR. This is a 3.5-, 2- and 1.5-fold increase compared to SequestHT, respectively, resulting on average in 2.5-fold more identified peptides per protein. We successfully demonstrated the fidelity of our new algorithm in four experiments: I) entrapment searches focusing on FDR-estimation, II) dilution experiments focusing on expected ratio distributions, III) comparisons with multiple search engines focusing on the overlap of identifications, IV) simulation experiments focusing on the deconvolution of chimeric spectra. Our new algorithm is compatible with older mass spectrometer generations, but profits disproportionally from the increased sensitivity of recent instruments and measurements using wider isolation windows. It substantially outperforms other search engines on data of different complexity such as body fluids and organisms from all kingdoms of life (Müller et al., 2020).

Conclusion: We present the first highly scalable, cloud-native, microservice-based and AI-powered search algorithm for the intensity-based deconvolution of chimeric spectra.

100321, https://doi.org/10.1016/j.mcpro.2022.100332

A.12
Computational challenges of identification and quantification algorithms in single-cell proteomic data.

Hannah Boekweg1

1Brigham Young University

The goal of proteomics is to identify and quantify the complete set of proteins in a biological sample. Single-cell proteomics specializes in the identification and quantitation of proteins for individual cells, often used to elucidate cellular heterogeneity. The significant reduction in ions introduced into the mass spectrometer for single-cell samples could impact the features of MS1 and MS2 spectra. As all identification and quantitation software tools have been developed on spectra from bulk samples and the associated ion-rich spectra, the potential changes in spectral features are of great interest. All identification and quantification algorithms need to be evaluated for single-cell data, and we investigated how peptide identification algorithms might underperform on single-cell data. We characterized the differences between single-cell spectra and bulk spectra by examining fundamental spectral features that are likely to affect peptide identification performance. All features show significant changes in single-cell spectra, including the loss of annotated fragment ions, blurring signal and background peaks due to diminishing ion intensity, and distinct fragmentation pattern, compared to bulk spectra. As each of these features is a foundational part of peptide identification algorithms, it is critical to adjust algorithms to compensate for these losses.

100321, https://doi.org/10.1016/j.mcpro.2022.100333