BIOGNOSYS
NEXT GENERATION PROTEOMICS

Biognosys AG
Wagistrasse 21
8952 Schlieren (Zurich)
Switzerland

t: +41 (0) 44 738 20 40
f: +41 (0) 44 738 20 49
info@biognosys.com
www.biognosyscom

Deep Learning Scores Improve Identification in Spectrum-centric Analysis of Proteomics Data

Maximilian J. Helf, Tejas Gandhi, Dariush Mollet, Tikira Temu, Marco Tognetti, Lukas Reiter

Introduction

Reliable generation of peptide-spectrum matches (PSMs) is a hallmark of mass spectrometry-based proteomics. Machine learning (ML) has been shown to significantly improve the quality of PSMs and is an integral part of Pulsar, the database search engine used by Biognosys software, including Spectronaut. Pulsar can perform spectrum-centric analysis of both DDA or DIA data using a protein database. It uses a wide range of quality scores as input for ML to discriminate between true and false PSMs. Some of these input scores are generated using deep learning (DL) models that were trained on large datasets and integrated into our pipeline. In this work, we investigated the impact of these DL scores on performance across a wide range of proteomics datasets.

Methods

Pulsar uses a multi-task deep neural network to predict the indexed retention time (iRT) and MS2 fragmentation of peptides. The delta iRT and spectral angle scores are then calculated based on how well the predicted and observed peptide properties match. The model was trained on 2.6 million PSMs to achieve a relative mean absolute error of 1.07% for iRT prediction and a median spectral angle score of 72% on test data (640`000 PSMs). DL scores from the model are calculated on-the-fly and used together with other scores to discriminate between true and false PSMs. We used Pulsar to search a broad range of datasets, with and without DL scores to compare identification numbers, empirical FDR and quantitative performance.

Preliminary data

For preliminary analysis, we analyzed 10 datasets from data-dependent acquisition (DDA) as well as data-independent acquisition (DIA), including MS data acquired on Thermo, Sciex and Bruker instruments. In modern proteomics, DDA data is most commonly used for isobaric labeling quantification (ILQ) or to generate libraries for DIA analysis. Spectrum-centric search of DIA data with Pulsar is the key first step of the directDIA pipeline that allows targeted analysis of DIA data without the need for an additional library.

Enabling DL scores led to modest improvement of identifications in DDA data amounting to a gain of up to 7% identifications in precursor and protein IDs after applying an FDR of 1% at PSM, peptide and protein group levels. When searching DIA data, DL scores had a markedly

Biognosys AG
Wagistrasse 21
8952 Schlieren (Zurich)
Switzerland

t: +41 (0) 44 738 20 40
f: +41 (0) 44 738 20 49
info@biognosys.com
www.biognosyscom

larger effect, with up to 15% and 10% gains in precursor and protein group IDs, respectively. This highlights the benefits of using deep learning models to classify true vs. false PSMs, especially for the analysis of the more complex and convoluted DIA data. We further investigated whether there were adverse effects to using DL scores in directDIA.

Using non-overlapping protein databases from two species, we determined that false discovery rate (FDR) was not significantly affected by the addition of DL scores. To assess quantitative performance, we analyzed data from a series of controlled quantitative experiments (CQEs) where proteomes from different species were mixed at known ratios between sample groups. We found that precision and accuracy of the analysis were comparable to using the analysis without DL scores despite the identification of an additional ~10% proteins. We conclude that the addition of DL scores is providing significant benefits while maintaining FDR and quantitative performance.

Novel aspect

Deep learning improves spectrum-centric analysis of both DDA and DIA data