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

Project Avalon: An Extended Machine Learning Framework for Spectronaut Dramatically Improves IDs for Short Gradient DIA Proteomics

Normand Overney, Oliver M. Bernhardt, Maximilian J. Helf, Tejas Gandhi, Lukas Reiter

Introduction

In recent years, sophisticated machine learning models have shown great potential in proteomics data analysis. Spectronaut, our DIA proteomics software, uses Linear Discriminant Analysis (LDA) for differentiating between true and false identifications. Here we evaluated several powerful classifiers by creating a new machine learning module in Spectronaut called Avalon. Although we could have used more complex machine learning algorithms, such as deep learning, we decided to tune classical ML classifiers by using a novel fitness function and genetic algorithms (GA) with the goal of improving identifications without compromising on reliability. In our experiments with the current state of Avalon, we show significant improvement in identification, especially for proteomics experiments with ultra-short gradients.

Methods

As part of Project Avalon, we evaluated eight different classifiers. The evaluation was done based on a fitness function which considers the number of identifications at a certain false discovery rate (FDR). Based on this, we decided to move forward with a gradient boosting machine (GBM) as our classifier. We then used GAs to tune our model and features using the same fitness function. The first GA would try to find the values for a select subset of tunable parameters for the GBM, while the second GA would independently find the best subset of features. Once we were confident in our results, we integrated Avalon into Spectronaut and tested it on short gradient dia-PASEF datasets by Meier et al.

Preliminary data

Recently the field has been experimenting with high-throughput DIA analysis by using very short LC gradients. Meier et al. in their recent publication used dia-PASEF acquisition ranging from 60 samples per day (~22 min per sample) to 200 samples per day (~7 min per sample). We benchmarked these datasets using Spectronaut 15 versus Spectronaut with Avalon because we expected a more powerful classifier to perform better with increasingly shorter gradients. For this benchmark, we settled on using a Gradient Boosting Machine.

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

Using Spectronaut with Avalon, we get an improvement at peptide precursor level of 20% with 60 SPD (93,135 precursors), 24% with 100 SPD (75,428 precursors), and 40% with 200 SPD (42,776 precursors). As expected, we see an increasing improvement with shorter gradients due to GBM performing better than LDA with increasingly complex data.

We also validated Spectronaut with Avalon by performing 2-species FDR tests based on three different datasets. Our empirical FDR calculation was comparable to Spectronaut 15 with LDA. Additionally, we also tested Avalon with data sets from three different vendors, different sample matrices, and longer gradients. We found a significant improvement in identification across the board (average 10-15% without short gradients).

Novel aspect

We developed a machine learning framework and tuned it with genetic algorithms to dramatically improve identifications in Spectronaut.