DeepDIA Demo: Training a New Model for Detectability Prediction

Training a new model for MS detectability prediction using data-dependent acquisition (DDA) data.

1. System Requirements

This demo has been tested on a workstation with Intel Xeon E5-2690 v3 CPU, 16 GB RAM, and Microsoft Windows Server 2016 Version 1607 (OS Build 14393.2430) operating system with the following softwares:

- Anaconda 4.2.0 (Python 3.5.2).
- Keras 2.2.4 and TensorFlow 1.11.
- Microsoft R Open 3.5.1.
- RStudio 1.1.447.
- Protein Digestion Simulator (https://omics.pnl.gov/software/protein-digestion-simulator).

A GPU card with Compute Unified Device Architecture (CUDA) is recommended, e.g. NVIDIA GeForce GTX 1050 Ti.

2. Demo Data

LC-MS/MS DDA data of HeLa and HEK-293 cells on Q Exactive HF are available at ProteomeXchange (http://proteomecentral.proteomexchange.org/) with the data set `PXD005573` . (Bruderer, R. et al. Mol. Cell. Proteomics 2017, 16, 2296-2309.)

- Fig4_HeLa-1m-HPRP-10perc_DDA_R01_T0.raw
- Fig4_HeLa-1m-HPRP-15perc_DDA_R01_T0.raw
- Fig4_HeLa-1m-HPRP-20perc_DDA_R01_T0.raw
- Fig4_HeLa-1m-HPRP-25perc_DDA_R01_T0.raw
- Fig4_HeLa-1m-HPRP-50perc_DDA_R01_T0.raw
- Fig4_HeLa-1m-HPRP-5perc_DDA_R01_T0.raw
- Fig4_HeLa-1m-HPRP-FT_DDA_R01_T0.raw
- Fig4_HeLa-1m_DDA_R01_T0.raw
- Fig4_HeLa-1m_DDA_R02_T0.raw
- Fig4_HeLa-1m_DDA_R03_T0.raw
- Fig4_HEK293-1m-HPRP-10perc_DDA_R01_T0.raw

- Fig4_HEK293-1m-HPRP-15perc_DDA_R01_T0.raw
- Fig4_HEK293-1m-HPRP-20perc_DDA_R01_T0.raw
- Fig4_HEK293-1m-HPRP-25perc_DDA_R01_T0.raw
- Fig4_HEK293-1m-HPRP-50perc_DDA_R01_T0.raw
- Fig4_HEK293-1m-HPRP-5perc_DDA_R01_T0.raw
- Fig4_HEK293-1m-HPRP-FT_DDA_R01_T0.raw
- Fig4_HEK293-1m_DDA_R01_T0.raw
- Fig4_HEK293-1m_DDA_R02_T0.raw
- Fig4_HEK293-1m_DDA_R03_T0.raw

SwissProt *Homo sapiens* database (FASTA) can be downloaded from UniProt (https://www.uniprot.org/). The FASTA file (2018-04 version, 20,301 entries)
has been deposited to ProteomeXchange via the iProX partner repository with the data set identifier `PXD014108/IPX0001628000`.

- swissprot_human_201804_validated.fasta

The saved project and exported results from SpectroMine are also available at ProteomeXchange/iProX with identifier `PXD014108/IPX0001628000`.

- HeLa_4h_DDA.psar.zip
- HeLa_4h_DDA.csv.zip
- HEK293_DDA.psar.zip
- HEK293_DDA.csv.zip

3. Prepare Training Data

In this demo, SpectroMine reports are used, which should be exported with the schema provided in the `misc/SpectroMine_Report_Schema` folder.

- PeptideReport.rs
- ProteinReport.rs

The file name of peptide report should end with `.PeptideReport.csv`, and that of protein report should end with `.ProteinReport.csv` e.g. `HEK293.PeptideReport.csv` and `HEK293.ProteinReport.csv`.

SpectroMine Manual is available at https://biognosys.com/shop/spectromine.

Start RStudio, ensure package `readr` has been installed.

```
install.packages("readr")
```

Open `deepdetect/R/init.R` and run the script by clicking `Source`.

```
source("{PATH_TO_CODE}/deepdetect/R/init.R")
```

Set the working directory to the reports and run `deepdetect/R/get_peptides_from_ProteinDigestionSimulator_result.R`.

```
setwd("{PATH_TO_DATA}")
source("{PATH_TO_CODE}/deepdetect/R/get_detectability_from_SpectroMine.R")
```

Single hits are excluded and only proteins with sequence coverage >= 25% are taken into considerasion. The protein list file are and the peptide detectability file are generated in the working directory.

- HEK293_excludeSingleHit_coverage25.proteinAccession.txt
- HEK293_excludeSingleHit_coverage25.detectability.csv

Open Windows PowerShell and run `deepdetect/Filter-Fasta.ps1`. A FASTA file containing the filtered proteins is generated in the working directory. Rename the filtered FASTA file.

```
{PATH_TO_CODE}/deepdetect/Filter-Fasta.ps1 swissprot_human_201804_validated.fasta HEK293_excludeSingleHit_coverage25.proteinAccession.txt
mv swissprot_human_201804_validated.filtered.fasta HEK293_excludeSingleHit_coverage25.fasta
```

Perform in silico digestion using Protein Digestion Simulator.
Tryptic (no Proline Rule) is selected as digestion enzyme. Digestion is performed with the following parameters:

- Max Miss Cleavages: 2
- Minimum Residue Count: 7
- Maximun Fragment Mass: 6000
- Minimun Fragment Mass: 0

DeepDIA only supports peptide sequences with standard amino acids (ACDEFGHIKLMNPQRSTVWY) and length <= 50.

Open `deepdetect/R/get_negative_peptides.R` and run the script by clicking `Source` .

```
source("{PATH_TO_CODE}/deepdetect/R/get_negative_peptides.R")
```

The negative peptide file is generated in the working directory.

- HEK293_excludeSingleHit_coverage25_negative.detectability.csv

Open `deepdetect/R/get_cleavage_window.R` and run the script by clicking `Source` .

```
source("{PATH_TO_CODE}/deepdetect/R/get_cleavage_window.R")
```

The training data are generated in the working directory.

- HEK293_excludeSingleHit_coverage25.detectability.csv
- HEK293_excludeSingleHit_coverage25_negative.detectability.csv

4. Train a MS Detectability Model

Run `deepdetect/py/train_hard_negative.py` to start training.

```
python {PATH_TO_CODE}/deepdetect/py/train_hard_negative.py
```

Expected run time depends on the number of peptides and the performance of the computer. In this demo, this command may take several hours to a day.

In the `training_{round}/models` folder, we find the trained model (with checkpoints during training).