Fast and accurate MS² peak intensity predictions for multiple fragmentation methods, instruments and labeling techniques

Poster · January 2019

DOI: 10.13140/RG.2.2.36500.55688

CITATIONS
0

READS
52

3 authors:

Ralf Gabriels
Ghent University
9 PUBLICATIONS **7** CITATIONS

SEE PROFILE

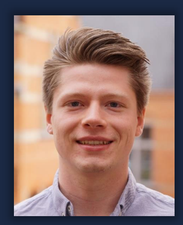
Lennart Martens
Ghent University
376 PUBLICATIONS **12,334** CITATIONS

SEE PROFILE

Sven Degroeve
Ghent University
64 PUBLICATIONS **4,357** CITATIONS

SEE PROFILE

FAST AND ACCURATE MS² PEAK INTENSITY PREDICTION FOR MULTIPLE FRAGMENTATION METHODS, INSTRUMENTS AND LABELING TECHNIQUES



Ralf Gabriels[1,2], Lennart Martens[1,2], Sven Degroeve[1,2]

1 VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium
2 Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

An important step in the analysis of high-throughput mass spectrometry-based proteomics is the correct identification of the peptide MS² fragmentation spectra obtained from a sample analysis. Due to incomplete understanding of the fragmentation process and unpredictable machine noise, matching MS² spectra with the correct peptide is far from trivial. We therefore developed MS²PIP: MS² Peak Intensity Prediction, a data-driven tool that accurately predicts the expected MS² spectrum for a given peptide.[1,2] Since its first publication, we have rebuilt MS²PIP from the ground up to be faster and more accurate. We also trained specific MS²PIP models for multiple specific cases: HCD and CID fragmentation, TripleTOF 5600+ instruments and iTRAQ- and TMT-labeled peptides. In each of these cases, the peak intensities are substantially influenced by the specific instrument or approach. Specific models therefore greatly improve the accuracy of MS²PIP.

Methods

To train and evaluate specific MS²PIP models, we downloaded and parsed a multitude of publicly available spectral libraries and experimental datasets. The size of the train-test datasets ranged from 129 000 to 1.6 million unique peptide spectra. The evaluation datasets contained between 3000 and 40 000 unique peptide spectra (Table 1). We can evaluate the models' performance by predicting MS² spectra present in the external evaluation datasets and comparing these predictions to their corresponding empirical spectra. This comparison is done by calculating the Pearson correlation coefficient of the two spectra, each normalized to their total-ion-current.

All scripts for training, testing and evaluating MS²PIP models are available on GitHub.com/CompOmics/MS2PIP_c.

Table 1: Train-test and evaluation datasets used to train specific MS²PIP models

Model	Use	Dataset	# Unique peptides
CID	Train-test	NIST CID	340 356
	Evaluation	NIST CID Yeast	92 609
HCD	Train-test	MassIVE-KB	1 623 712
	Evaluation	PXD008034	35 269
iTRAQ	Train-test	NIST iTRAQ	704 041
	Evaluation	PXD001189	41 502
iTRAQ phospho	Train-test	NIST iTRAQ phospho	183 383
	Evaluation	PXD001189	9 088
TMT	Train-test	Peng Lab TMT Spectral Library	1 185 547
	Evaluation	PXD009495	36 137
TTOF5600	Train-test	PXD000954	215 713
	Evaluation	PXD001587	15 111

Results

The median Pearson correlations between empirical spectra and spectra predicted with the corresponding specific models are consistently higher then when we apply other models to the same data set. Only the specific TTOF5600 model is essentially matched by the HCD model when predicting TTOF5600 spectra. Predictions from the correct models yield median Pearson correlations higher then 0.90, except for the TTOF5600 and the iTRAQ phospho models, which have median Pearson correlations of 0.74 and 0.84, respectively (Figure 1).

It is also noteworthy that models for labeling techniques perform similarly on all datasets, indicating that TMT and iTRAQ labels affect the fragmentation pattern in a comparable fashion (Figure 1).

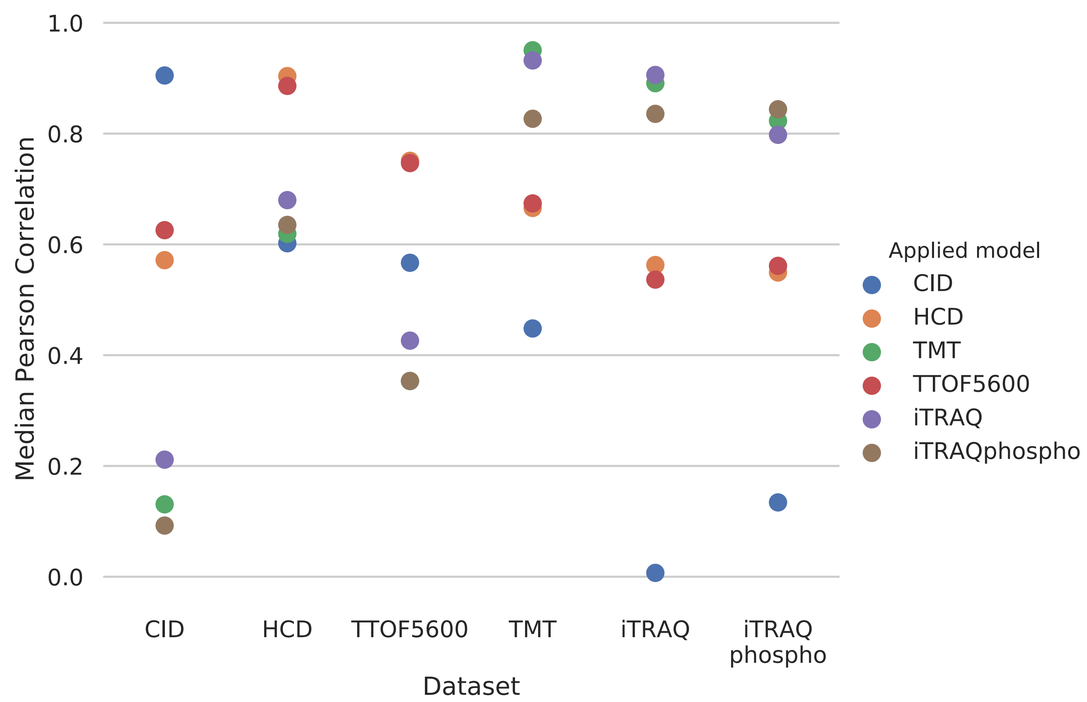


Figure 1: Median Pearson correlations for all specific models, applied to all evaluation datasets

Discussion and conclusions

These results confirm that training specific peak intensity prediction models for specific cases substantially improves the predictions. This can also be visually confirmed when comparing predictions from the HCD and TMT model, respectively, with an empirical TMT spectrum (Figure 2).

MS²PIP has already been used for creating proteome-wide spectral libraries for search engines (including Data Independent Acquisition), for selecting discriminative transitions for targeted proteomics[3,4], and for validating interesting peptide identifications (e.g. biomarkers) [5,6].

These new models extend the applicability of MS²PIP even further, allowing it to be applied to specific fragmentation methods, instruments, or labeling techniques.

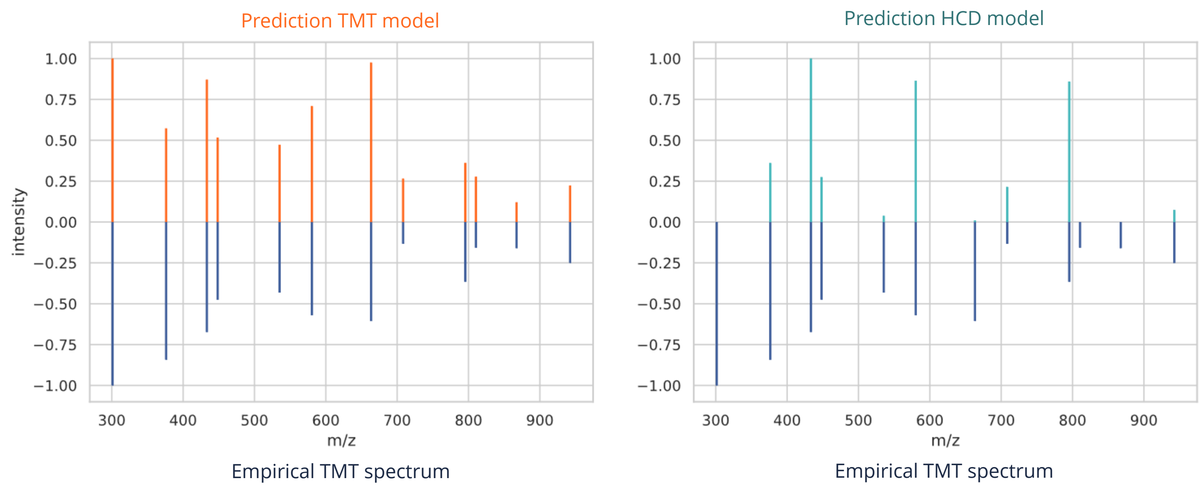


Figure 2: Spectra predicted by MS²PIP TMT model (top left) and HCD model (top right) compared to an empirical spectrum of a TMT-labelled peptide (bottom left and right).

TRY OUT MS²PIP YOURSELF! GO TO OUR USER-FRIENDLY WEBSERVER AND DEVELOPER-FRIENDLY RESTFUL-API AT IOMICS.UGENT.BE/MS2PIP

1. Degroeve,S. et al. (2015) Nucleic Acids Res., 43, W326–W330. doi:10.1093/nar/gkv542
2. Degroeve,S. et al. (2013) Bioinformatics, 29, 3199–203. doi:10.1093/bioinformatics/btt544
3. Albrethsen,J. et al. (2018) Clin. Chem. Lab. Med., 56, 1913–1920. doi:10.1515/cclm-2018-0171
4. Mesuere,B. et al. (2016) Proteomics, 16, 2313–2318. doi:10.1002/pmic.201600023
5. Budamgunta,H. et al. (2018) Proteomics, 18, 1700218. doi:10.1002/pmic.201700218
6. Willems,P. et al. (2017) Mol. Cell. Proteomics, 16, 1064–1080. doi:10.1074/mcp.M116.066662