SUPPLEMENTARY MATERIAL 1

DeepMSPeptide: peptide detectability prediction using deep learning.

Guillermo Serrano[1,#], Elizabeth Guruceaga[1,2,#] and Victor Segura[1,2,*]

[1]Bioinformatics Platform, Center for Applied Medical Research, University of Navarra, Pamplona, 31008, Spain. [2]IdiSNA, Navarra Institute for Health Research, Pamplona, 31008, Spain.

[*]Corresponding author: Victor Segura, Bioinformatics Platform, Center for Applied Medical Research, Avenida Pío XII 55, E31008 Pamplona, Spain. Tel: +34 948194700, e-mail: vsegura@unav.es

GPMDB benchmark

The GPMDB dataset for Homo sapiens was downloaded on 15 January 2019 from its website (http://peptides.thegpm.org/~/peptides_by_species/). Therefore modern datasets have been considered as well as old datasets that have been approved by the GPMDB quality control AI. A detailed description of GPMDB data sources is available in https://wiki.thegpm.org/wiki/GPMDB_Data_Sources, including citations to 2018 datasets of published papers from 2005 to 2019. Consequently, we are predicting the intrinsic detectability of a peptide considering GPMDB information about the number of detections per each peptide in any experiment, and GPMDB includes experiments that use different mass spectrometers.

‘Missing proteins’ benchmark

As an additional benchmarking we used the MS evidence for the human proteome provided by the HPP project (Legrain,P. et al., 2011) and extracted from the neXtProt database (Gaudet,P. et al., 2017). We divided the peptides of this database, also contained in GPMDB but independent to the training and test sets of peptides, into 3 non-overlapping groups: (1) proteins with MS evidence (PE1) in the current release (2019-01-11); (2) PE1 proteins in the current release without MS evidence (*missing proteins*) at the beginning of the HPP project (2011-08-23); (3) missing proteins (MPs) in the current release. We are assuming that the peptides of PE1 proteins are more detectable than the peptides from the missing proteins (MPs) which currently lack of experimental validation. However, it is also true that some PE1 proteins have low detectable peptides and some MPs could have some high detectable peptides. For a balanced comparison we only considered 8000 unique peptides randomly selected from each group as shown in Table 1.

Table 1. Number of proteins and peptides of the defined groups for the comparison of peptide detectability classifiers.

	PE1 proteins	Detected MPs	Current MPs
Peptides	1338772	280271	95890
Proteins	13305	3132	2127
Unique peptides[1]	1257768	248924	74632
Proteins with unique peptides	13222	3123	2050
GPMDB observation filter[2]	69717	9524	72942
Corresponding proteins	5941	1436	2048
Peptides independent to training and testing datasets	31018	9423	72942
Corresponding proteins	5640	1433	2048
Selected peptides	8000	8000	8000
Selected proteins	3556	1359	1779

[1] Only unique peptides were considered to avoid shared peptides among proteins with different protein evidences.

[2] A filtering process using the number of MS/MS observations in GPMDB was used to define a set of peptides from PE1 proteins and Detected MPs with a high number of identifications and a set of peptides from MPs that were not present in GPMDB.

The results obtained using the Random Forest classifier and the 1D-2C-CNN showed that the deep learning algorithm was able to provide a more robust prediction of detectability (~1) for the peptides of PE1 proteins and non-detectability (~0) for the peptides of the MPs (Figure 1). Interestingly, the classifiers considered detectable the set of peptides corresponding to those MPs that in the current neXtProt release are classified as PE1 based on experiments performed to validate their MS evidences.

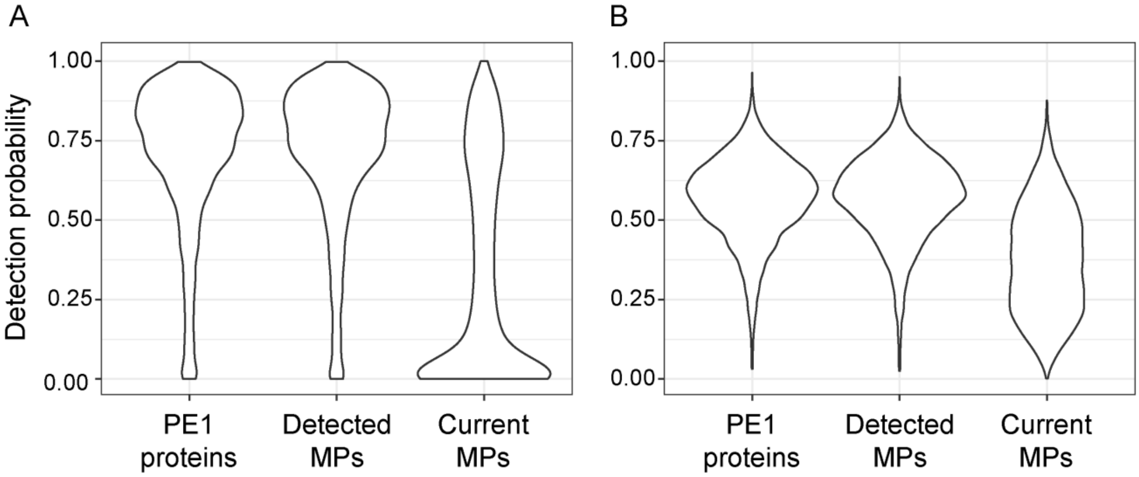

Figure 1. Benchmarking of the 1D-2C-CNN (A) and Random Forest classifier (B). The results of the peptide detectability prediction for the peptides with different protein evidences in neXtProt database are shown.

References

Legrain,P. et al. (2011) The human proteome project: Current state and future direction. Mol. Cell Proteomics, 10(7), M111.009993.

Gaudet,P. et al. (2017) The neXtProt knowledgebase on human proteins: 2017 update. Nucleic Acids Res., 45(Database issue), D177–D182.