Deep Learning based Peptide Detectability in Shotgun Proteomics

*Juho Son[‡,1], Seungjin Na[1] and Eunok Paek[*1]*

[1]Department of Computer Science, Hanyang University, Seoul, Republic of Korea.

INTRODUCTION

Peptide detectability is defined as a probability of being observed in shotgun proteomics, that includes both the possibility of the presence of peptide digested in protein and whether the peptide is ionized and detected in LC-MS/MS. Detectability can be used to reduce the size of the sequence database to be used for peptide identification by database search, or can be useful for protein inference. [1] Several computational approaches such as AP3 [2], DeepMSpeptide [3] and PepFormer [4] have been proposed to predict peptide detectability via machine learning. DeepMSpeptide and PepFormer predict the detectability through sequence embedding only by looking at the peptide sequence. In contrast, AP3 focuses on the digestive process based on the protein's physicochemical properties and digestibility rather than sequential embedding. Here we propose an end-to-end network model that can capture the digestion process by considering enzymatic sites as input to enhance the detectability prediction performance.

METHODS

Dataset

We used the massIVE-KB [5] archive to obtain peptides with confirmed tryptic digestion sites, and proteins with a sequence coverage of 0.5 or higher to confirm their existence and amenability to mass spectrometry. Among the peptides in massIVE-KB, the detected peptides with a spectral count of 2 or more were used as positive training data, and undetected peptides were used as negative training data while allowing up to two missed cleavages in identified proteins. All peptides were fully tryptic and their lengths were limited to 7-30. As an input, a total of 15-mer was used, 7-mer around K and R, so that the protein digestion site is positioned in the middle. AAindex [6] was adopted to reflect the physicochemical properties of peptides. Table 1 and 2 show our datasets.

Multi input end-to-end network

We propose a multi-input end-to-end model with peptide sequences, tryptic site sequences, and physicochemical properties as input. The network first receives five inputs consisting of label encoded sequences and physicochemical properties. Sequence embedding dimensions are 32 and 16, respectively, and the embedded vector is used for bidirectional LSTM with 32 units. Finally, each input is concatenated with a fully connected layer of 80 dimensions. We used 300 for epoch, 128 for batch size, and 1e-4 for learning rate. Loss used binary cross entropy. Figure 1 shows our workflows.

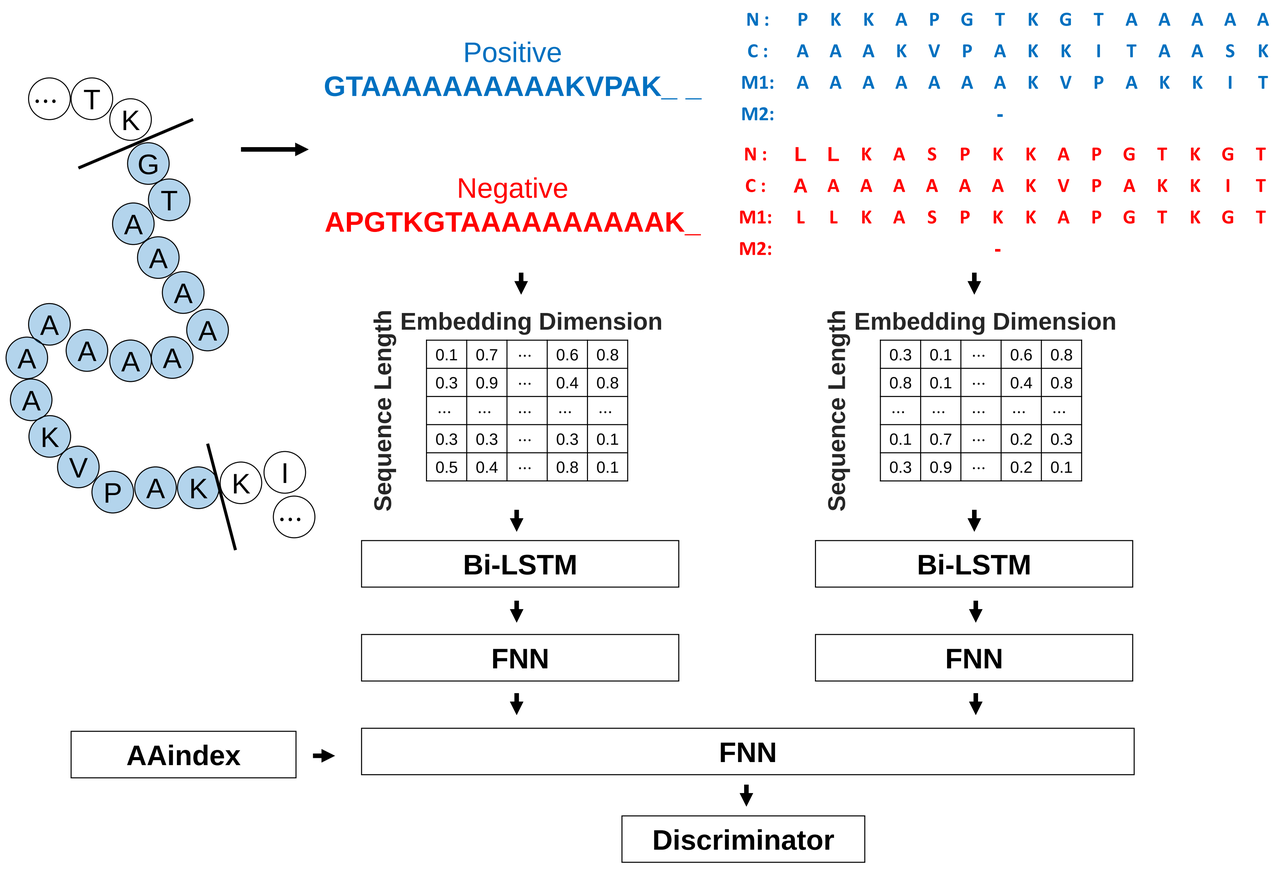

Figure 1. Workflow of our model. Peptide sequence, tryptic sites of N, C terminal, and miss cleavage, and AAindex are input respectively.

data set	proteins idedtified	peptides identified	proteins after filtration	peptides after filtration
massIVE-KB	19,291	506,605	11,707	485,857

Table 1. number of identified proteins and peptides in massIVE-KB dataset. For protein existence, we filter out proteins with sequence coverage 0.5.

data set	number of peptides
train	215,352
validation	53,838
test(holdout)	67,298

Table 2. number of peptides in train, validation, and test set. In the peptides after filtration, we used only half of the data for comparison with AP3, which requires separate digestibility learning.

RESULTS

Table 3 describes the necessity of a multi-input network. AAAAAAAKVPAKKIT, a 15-mer tryptic site, corresponds to the area that should be digested as a C terminal tryptic site for red and yellow peptides. However, in the case of peptide identified with this 15-mer site as missed cleavage, such as green and blue, it should not be digested. As can be seen in this example, even the same 15mer tryptic site may or may not be digested depending on the peptide sequence. Therefore, in order to capture the digestion process well and use it to predict peptide detectability, the model should be constructed with multiple inputs rather than learning digestibility separately.

Table4 shows that our model clearly performs better than other existing models.

Peptide	Tryptic site of N-terminal	Tryptic site of C-terminal	Tryptic site 1 of miss cleavage	Tryptic site 2 of miss cleavage
K.KAPGTKGTAAAAAAAAAAK.V	ALLKASPKKAPGTKG	**AAAAAAAKVPAKKIT**	LLKASPKKAPGTKGT	PKKAPGTKGTAAAAA
K.GTAAAAAAAAAAK.V	PKKAPGTKGTAAAAA	**AAAAAAAKVPAKKIT**	-	-
K.GTAAAAAAAAAAKVPAK.K	PKKAPGTKGTAAAAA	AAAKVPAKKITAASK	**AAAAAAAKVPAKKIT**	-
K.GTAAAAAAAAAAKVPAKK.I	PKKAPGTKGTAAAAA	AAKVPAKKITAASKK	**AAAAAAAKVPAKKIT**	AAAKVPAKKITAASK

	P	K	K	A	P	G	T	K	G	T	A	A	A	A	A	A	A	A	A	A	K	V	P	A	K	K	I	T
Spectral count	0	14	3	0	109	0	2	7	126	92	92	92	92	92	92	92	92	92	92	93	185	0	0	0	28	51	0	0
miss count	44	30	33	40	149	149	151	144	178	178	178	178	178	178	178	178	178	178	178	179	86	103	96	96	68	25	25	25

Table 3. Example of sequence of peptides and tryptic sites. Spectral counts refer to the number of times the spectrum of peptides that do not contain miss cleavage is observed at each amino acid position. On the contrary, miss count is the case of including miss cleavage.

Models	AUC	ACC
AP3	0.875	0.790
DeepMSpeptide	0.871	0.809
PepFormer	0.906	0.832
ours	0.917	0.836

Table 4. Performance of State-of-the-Art Models for peptide detectability prediction.

CONCLUSIONS

In this study, we presented a multi-input end-to-end network. For the first time, considering that the protein digestion process was not well reflected past, the sequence of peptides and tryptic sites was received as input in one network. Our results showed that the model effectively predicts the peptide detectability.

REFERENCES

(1) Li, Y. F., Arnold, R. J., Tang, H., & Radivojac, P. (2010). The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics. *Journal of proteome research*, *9*(12), 6288-6297.
(2) Gao, Z., Chang, C., Yang, J., Zhu, Y., & Fu, Y. (2019). AP3: An Advanced Proteotypic Peptide Predictor for Targeted Proteomics by Incorporating Peptide Digestibility. *Analytical chemistry*, *91*(13), 8705-8711.
(3) Serrano, G., Guruceaga, E., & Segura, V. (2020). DeepMSPeptide: peptide detectability prediction using deep learning. *Bioinformatics*, *36*(4), 1279-1280.
(4) Cheng, H., Rao, B., Liu, L., Cui, L., Xiao, G., Su, R., & Wei, L. (2021). PepFormer: End-to-End Transformer-Based Siamese Network to Predict and Enhance Peptide Detectability Based on Sequence Only. *Analytical Chemistry*, *93*(16), 6481-6490.
(5) Wang, M., Wang, J., Carver, J., Pullman, B. S., Cha, S. W., & Bandeira, N. (2018). Assembling the community-scale discoverable human proteome. *Cell systems*, *7*(4), 412-421.
(6) Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. (2007). AAindex: amino acid index database, progress report 2008. Nucleic acids research, 36(suppl_1), D202-D205.

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-01373, Artificial Intelligence Graduate School Program(Hanyang University), National Research Foundation of Korea (NRF) grants (2019M3E5D3073568), and by the BK21 FOUR (Fostering Outstanding Universities for Research) project of the National Research Foundation of Korea Grant.