DbDeep : Peptide Detectability Prediction by Deep Learning

Juho Son[1], Seungjin Na[2] and Eunok Paek[1,2]

[1]Department of Computer Science, Hanyang University, Seoul 04763, Republic of Korea,
[2]Institute for Artificial Intelligence Research, Hanyang University, Seoul 04763, Republic of Korea

INTRODUCTION

Peptide detectability is helpful in various applications of mass spectrometry (MS)-based proteomics, in particular targeted proteomics. 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], PepFormer [4] and CapsNet+CBAM [5] have been proposed to predict peptide detectability via machine learning. DeepMSpeptide, PepFormer and CapsNet+CBAM predict the detectability based on the peptide sequence only. AP3 focused on peptide digestibility and physicochemical properties. Here we propose an end-to-end network model that predicts peptide detectability by embedding the peptide sequence and the cleavage site contexts at the same time.

METHODS

Dataset

We used the MassIVE-KB [6] spectral library to obtain peptides with confirmed tryptic digestion sites, and proteins with a sequence coverage of 50% or higher. 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. A peptide could have three type of cleavage sites 1) N-terminal, 2) C-terminal, and 3) missed cleavage sites. Each cleavage site is represented as a 15-mer, with the cleavage site amino acid, Lys (K) or Arg (R) for trypsin, and its 7 flanking amino acids on each side. Table 1 and 2 show our datasets.

Spectral library	#Observed proteins	#Observed peptides	#Proteins after filtration	#Peptides after filtration
MassIVE-KB	19,300	1,075,832	12,042	949,623

Table 1. The numbers of identified proteins and peptides in MassIVE-KB library. To select regularly digested peptides from confidently identified proteins, we retained 408,208 fully tryptic peptides (including up to two missed cleavages) from 12,042 proteins with a sequence coverage of 50% or higher.

Multi input data-centric network

We propose a multi-input end-to-end model with peptide sequences and tryptic site sequences as input. The network first takes five inputs consisting of label encoded sequences. 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 200 for epoch, 128 for batch size, and 1e-4 for learning rate. Loss used binary cross entropy. Figure 1 shows the architecture.

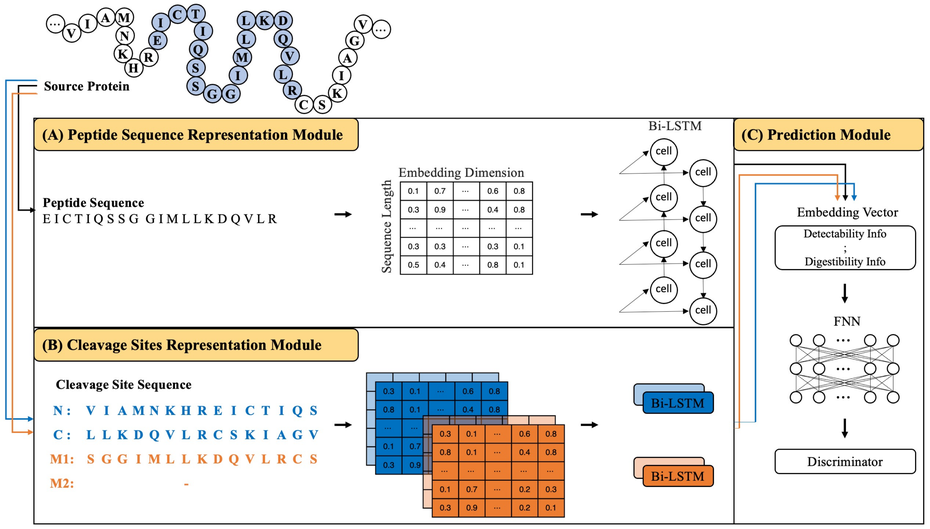

Figure 1. Model architecture of DbDeep. DbDeep takes a peptide sequence, N-terminal, C-terminal, and missed cleavage site sequences and outputs the detectability of the input peptide.

data set	number of peptides
train	536,492
validation	134,124
test(holdout)	170,367

Table 2. The number of peptides in train, validation, and test set.

Models	ACC	AUC
Ours	0.8412	0.9212
AP3	0.7855	0.8702
DeepMSpeptide	0.8165	0.8973
PepFormer	0.8351	0.9132
CapsNet+CBAM	0.8268	0.9067

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

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.

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 shown in this example, even the same 15mer tryptic site may or may not be digested depending on the peptide sequence. Therefore, in order to take into account the digestion process in predicting peptide detectability, the model should be constructed with multiple inputs rather than learning digestibility separately.

Peptide	Tryptic site at N-terminus	Tryptic site at C-terminus	Tryptic site 1 of missed cleavage	Tryptic site 2 of missed 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	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
Missed 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 mean the number of peptide spectra that do not contain missed cleavage, while missed count is the case of including missed cleavage.

CONCLUSIONS

In this study, we presented a novel end-to-end LSTM network model that combines the contexts of peptides and cleavage sites (by protease) for peptide detectability prediction. Utilizing the cleavage site contexts could improve the prediction performance because the digestion significantly affected the peptide detection in MS/MS. Table4 shows that DbDeep outperformed the existing methods.

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) Yu, M., Duan, Y., Li, Z., & Zhang, Y. (2021). Prediction of Peptide Detectability Based on CapsNet and Convolutional Block Attention Module. International journal of molecular sciences, 22(21), 12080.
(6) 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.