Construction of a predictive convolutional neural network model for antimicrobial peptide sequences

Yu jeoung Shim[1], Lee Byungjo[1], Wonhee Jang[1*]

[1]*Department of Life sciences, Dongguk University, Seoul 100-713, Republic of Korea*

Abstract

Nowadays, antibiotic resistance has become a major medical problem caused by overuse of antibiotics. Antimicrobial peptides (AMPs) are spotlighted to overcome bacterial drug-resistance. AMP is short cationic peptides that can directly interact with bacterial components, which cause membrane disruption. However, AMP not causes antibiotic resistance with not change bacterial membrane completely. Also, AMP can be obtained from animal transcriptome that is a good source of bio-compound with various properties. We constructed supervised deep learning models using convolutional neural network (CNN) method. We collected sequences of AMPs and non-AMPs from the public database for model training. We trained the models with the training dataset and selected the model with the highest accuracy with the test dataset. Putative AMP peptides were screened from the spider's transcriptome sequence data using the selected CNN model. In our study, we predicted novel AMPs from the transcriptome of spider using the best performing CNN model. The results indicated the utilization of deep learning to the discovery of new antibiotics against drug-resistance strains.

Convolutional Neural Network Model

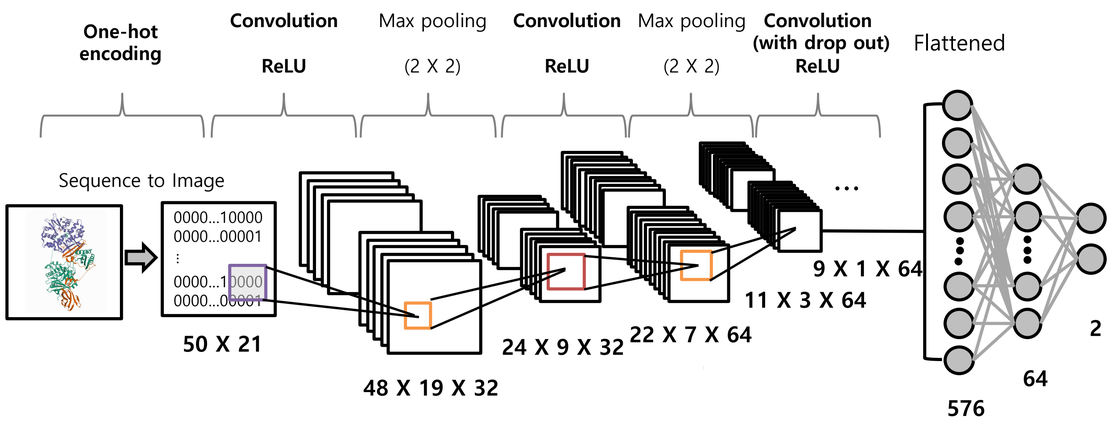

Figure 1. The model for screening AMP. The model consisted of 3 of 2D convolution layers with ReLU activation function and 2 max pooling layers and a dropout layer.

Dataset Information

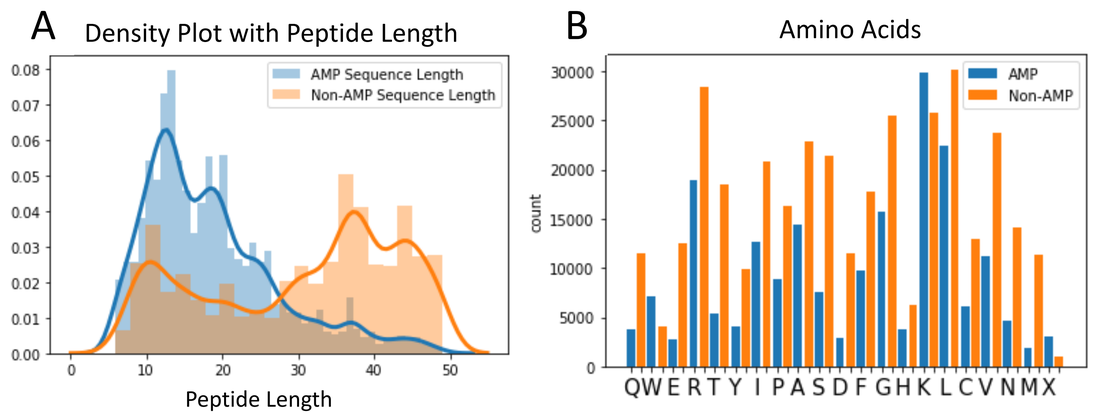

Figure 2. As the data used for training, 10954 AMPs from Database of Antimicrobial Activity and Structure of Peptides (DBAASP) and 10954 non-AMPs from UniProt with a length of more 5 and less 50 were used. **A.** Histogram of peptide length according to the presence of antibacterial function. **B.** Bar plot of the number of each amino acid that composed peptides.

Training and Validation

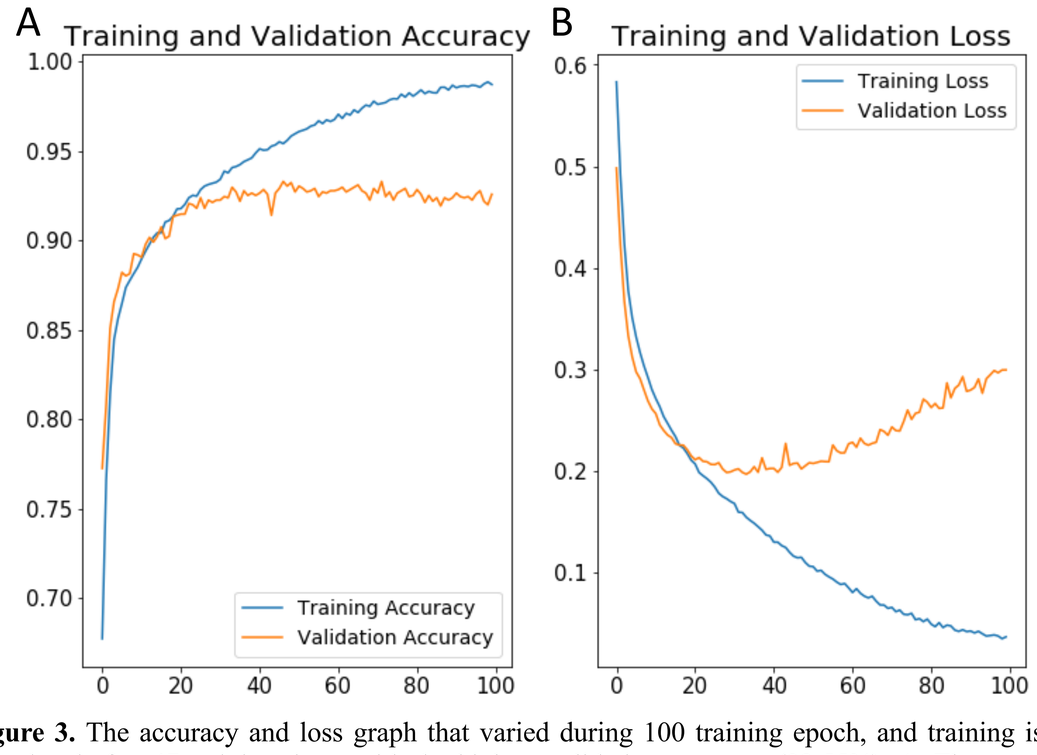

Figure 3. The accuracy and loss graph that varied during 100 training epoch, and training is completed after 47 training times with the highest validation accuracy (93.55%). **A.** The trend of training and validation accuracy during 100 training epoch. **B.** The trend of training and validation loss during 100 training epoch.

Model Test

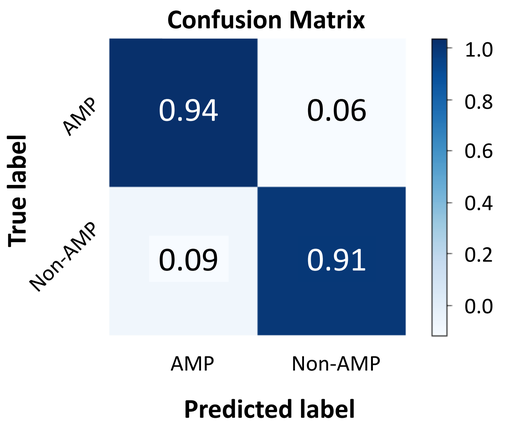

Table 1. Classification matrix.

	Precision	Recall	F1-score	Accuracy
AMP	0.91	0.94	0.93	92.59%
Non-AMP	0.94	0.91	0.92	

Figure 4. The confusion matrix of classified AMP and non-AMP test set. The trained CNN model achieved an accuracy of 92.59% in classification of AMP.

Screening Putative AMP Workflow

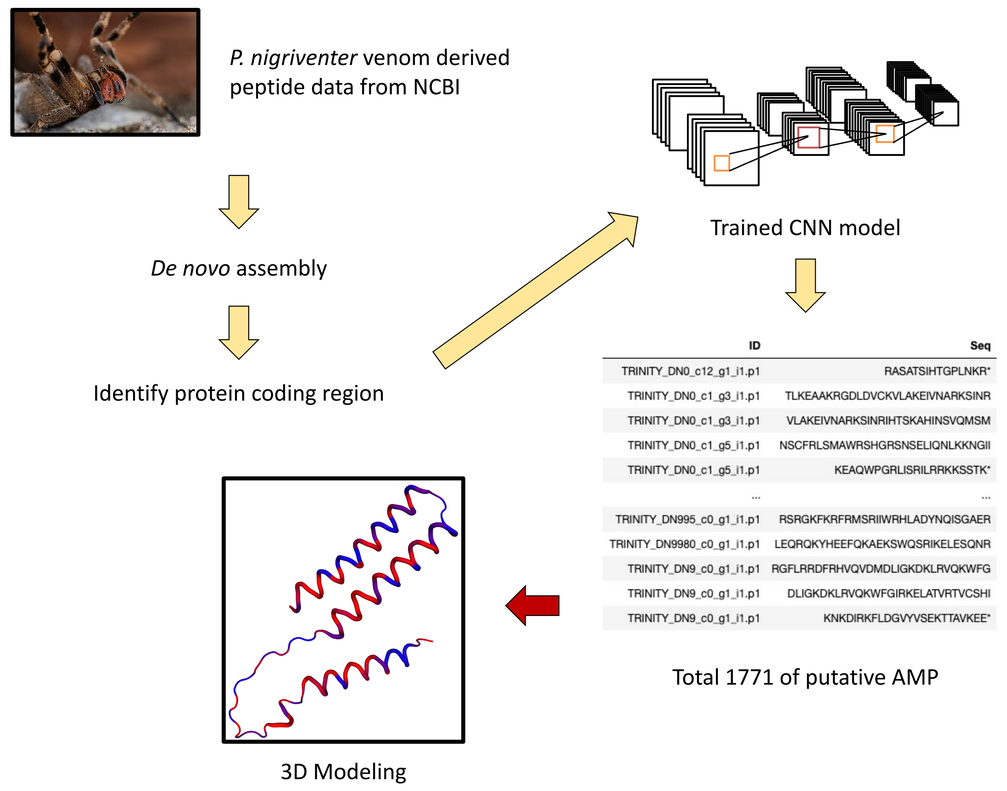

Figure 5. The Workflow of the AMP screening among spider venom peptides. Using the *P. nigriventer* spider peptide data registered in the National Center of Biotechnology Information (NCBI) database, a total of 1771 peptide sequences were predicted to have AMP function. In order to find out whether it has an alpha helix structure, which is a characteristic of AMP, 3D modeling was performed on putative sequences.

Conclusion

We trained, validated and tested a CNN model using non-AMP data collected in UniProt and AMP data found in DBAASP to create an optimal model for screening AMP. Using this learned model, we predicted whether peptides of length 50 or less derived from *P. nigriventer* spider venom from NCBI database could function as AMP. As a result, a list was made that predicts that 1771 of the peptides of 30335 will be capable of anti-bacterial function. However, there are limitations to our study. This model can only screen peptide sequences less than 50 in length. Therefore, the spider-derived toxic peptide obtained from NCBI was predicted by cutting into 50 peptide units. However, there is a possibility that AMP cannot be predicted if there is AMP in the cut part. We will overcome this limitation through future research.

Future Work

We intend to further update this predictive model with future research to develop a model that screens the part of the spider's venom-derived peptide that functions as AMP even in peptides longer than 50 amino acids in length. By recognizing long peptide segments in sequence with our trained model, we will be able to predict whether a long peptide is AMP or not without using long AMP data for training. The research has important implications for screening new AMPs and is expected to play a role in triggering faster AMP validation studies.

Acknowledgement

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2021-2020-0-01789) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)