

Predicting peptide bioactivity with long short-term memory: a compact model for sequence-based classification

¹Sy, I.C.B.C and ^{2,*}Sarza, R.M.O.

¹*Electronics and Electrical Engineering Institute, College of Engineering, University of the Philippines, Diliman, Quezon City, Philippines*

²*Department of Chemistry, Faculty of Science, National University of Singapore, Singapore*

Article history:

Received: 24 April 2024

Received in revised form: 20 January 2025

Accepted: 13 March 2025

Available Online: 19 March 2025

Keywords:

Bioactive peptides,
Predictive modeling,
Long short-term memory

DOI:

[https://doi.org/10.26656/fr.2017.8\(S8\).12](https://doi.org/10.26656/fr.2017.8(S8).12)

Abstract

Bioactive peptides (BPs) are short-chain amino acids that have been shown to have several health benefits. Laboratory-based identification of BPs is a time-consuming and labor-intensive process, hence multiple in-silico methods have been developed and used to determine BPs from various sources. In this study, a model, based on a long short-term memory (LSTM) binary classification model, was created and trained on existing bioactive peptides from literature and available databases to determine if an input peptide is bioactive or not. Peptide sequences used were from plant and animal sources. In addition, a set of non-bioactive peptides was used as the null set. The study used word encoding on each sequence wherein each sequence was preprocessed to have spaces in between each amino acid before being tokenized. Results showed an 89.13% validation Accuracy and a 0.8174 f-measure value for the model that utilized the dataset containing all peptides used. Model performance can be improved by adding to the dataset and trying different encoding methods, which retain more information. The constructed model and its prediction can be used as a baseline for further studies in the predictive determination of bioactive peptide sequences from food products.

1. Introduction

Bioactive peptides (BPs) are short-chain amino acids that have shown activity in terms of providing health benefits to the consumer. BPs are produced during the enzymatic breakdown that proteins undergo during digestion, or in other cases, food processing (Daliri *et al.*, 2017; Toldra *et al.*, 2018). There is an increasing interest in these macromolecules in different fields and industries due to the fact that they can provide several benefits not just in terms of health, but also in terms of food safety and quality. Several studies have shown that bioactive peptides can provide antihypertensive, antioxidative, and immunomodulatory health benefits to name a few. Likewise, in terms of food product quality, there are bioactive peptides that have shown antimicrobial properties (Udenigwe and Aluko, 2012; Daliri *et al.*, 2017; Sanchez and Vasquez, 2017). This interest resulted in numerous laboratory-based studies focusing on identifying new bioactive peptide sequences and deducing their functions. However, one of the major drawbacks of these studies is the labor-intensive and time-consuming process of experimentation (Li *et al.*, 2022). To address this, if an initial prediction can be

made on the peptide sequence, we can reduce the resources needed.

Machine learning is a field that uses different computational algorithms designed to copy the human cognitive process. One approach to machine learning is neural networks wherein several layers of interconnected algorithms are used to analyze inputs and generate outputs similar to how the human brain processes data (Naga and Murphy, 2015; Ma *et al.*, 2022). Different neural networks have been developed as classification or predictive models in various fields, and some of these have already been used in predicting bioactive peptide sequences. An example of this is the bidirectional recurrent neural network (BRNN) model that uses sequential data used in predict bioactive peptide sequences from larger protein sequences and the protein's predicted structural features (Schuster and Paliwal, 1997; Mooney *et al.*, 2013). A disadvantage of this model is the need to have the original protein sequence available which may not be possible for pre-digested samples. Another model used in bioactive peptide research is the convolutional neural network (CNN). CNN uses kernels that allow it to extract features

*Corresponding author.

Email: riann.sarza@u.nus.edu

on its own. The model has been used in predicting the anti-hypertensive activity of certain peptides given several properties of the peptide as inputs (Shi and Zhang, 2022). These applications are considered examples of bioinformatics. On the other side of this, molecular docking uses simulations and a deeper understanding of the molecular structure, its physicochemical properties, and its interactions both inter and intramolecular to make a prediction (Vidal-Limon *et al.*, 2022). What is common between the two different methods is that both acknowledge that multiple variables are in play to make a prediction.

Recurrent neural networks (RNN) are mostly used in time series, also known as sequential data. RNN has a node which stores the data from the previous state to use in current predictions. Although RNN has problems in creating long-term dependencies, in BPs' case, it might not recognize an amino acid 20 positions before the current state. Long short-term memory (LSTM), shown in Figure 1, is a type of RNN which can retain more information. It uses additional gates that tell the node if it should forget, store, or output the information it has, which alleviates the weights being too small or too large (Hochreiter and Schmidhuber, 1997; Guo *et al.*, 2020). These gates have the function of choosing what information is stored, maintained, and removed from each cell.

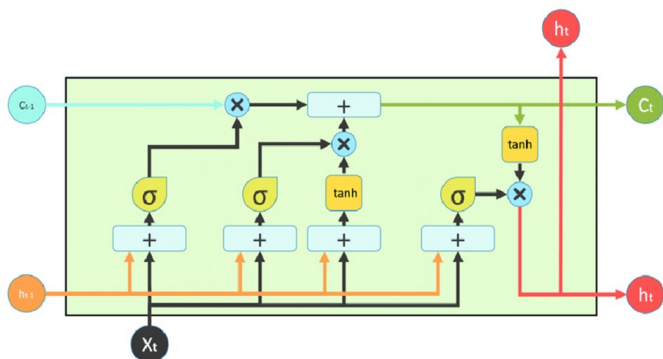


Figure 1. LSTM network structure. Adapted from Guo *et al.* (2020).

This work aimed to address the lack of predictive models that only use the peptide sequence to determine whether or not it has bioactivity. It aimed to determine if a model can be built with fewer variables unlike common bioinformatics or molecular docking approaches. LSTM is most fit in this application due to the possible dependencies between longer chains of amino acids, with only the sequence data being the input of the model. It is important to form these relationships, to determine how accurate a prediction can be using minimal information to maximize performance.

2. Materials and methods

Figure 2 shows the process of creating the dataset,

model, and metrics used in this study. The peptide sequences utilized were all obtained from existing literature and came from plant and animal food sources. Peptide sequences were obtained from the BIOPEP-UWM database (Minkiewicz *et al.*, 2019). Non-bioactive protein sequences were obtained using the protein sequences available within the database and fragmenting via the enzymatic digestion function of BIOPEP-UWM. Bioactive peptide sequences on the other hand were obtained using both the BIOPEP-UWM database and those pre-existing in published literature. Fragments are then labeled as “nonbioactive” or “bioactive” accordingly. A total of 4672 bioactive peptides and 1123 non-bioactive were extracted, 2731 of which were of lengths 5 or less and of that, 2036 were bioactive. Of the 3064 sequences with lengths of more than 5, 2637 of which are bioactive.

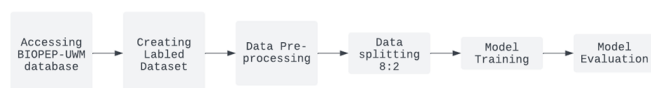


Figure 2. AI workflow used in the study.

The data were separated into three (3) sets: peptide chains with 5 or fewer amino acids (s1), more than 5 (s2), and a set of all peptides (s3). The peptide sequences were then run in an Excel function to add spaces in between each amino acid and saved as csv. The labeled dataset created was imported to MATLAB and separated into training and validation datasets at a 8:2 ratio, respectively. Datasets then underwent the MATLAB function *tokenizedDocument* to treat each of the data as a document. *wordEncoding* is then applied which maps each amino acid to numeric indices, then converting the documents into sequences with lengths five (5) for the s1 and twenty (20) for s2 and s3 to prepare as input for the model. The way the model works is similar to how documents are encoded, where certain keywords are the triggers, in this case, bioactive or not.

The model has six (6) layers consisting of a sequence input layer, a word embedding layer with *embeddingDimension* of 50, LSTM layer with 10 hidden units, a Fully connected layer, a softmax layer, and a Classification Output layer. This approach is often used to process text documents, in this case, each amino acid is being treated as a word in a sentence or document. The position of each amino acid is not considered. The model is then trained using the training settings in Table 1. The training dataset was used to train the model directly, while the validation dataset was used every 50 iterations to evaluate performance during training.

The model was evaluated using the f-measure and Accuracy, derived from the confusion matrix from the validation split. Accuracy is the amount of correctly

classified peptides over the total number of peptides. F-measure is the weighted average of precision and recall. Accuracy is highest at 100% while f-measure is highest at 1.

Table 1. Setting used to train the predictive model.

Solver	adam
MiniBatchSize	128
GradientThreshold	2
Shuffle	every-epoch

3. Results and discussion

Datasets s1 and s3 were trained to 200 epochs, while s2 was trained up to 50 epochs. S1 and s3 accuracy stabilized at around 50 epochs and maintained that accuracy. The s2 dataset was only trained until 50 epochs, after that point, accuracy started to drop due to possible overfitting due to the smaller size of the s2 dataset. The macro average accuracy and macro average f-measure of the model under each dataset are seen in Figure 3 where s3 achieved the highest Accuracy and f-measure at 89.13% and 0.8174 respectively among the 3 datasets. It could have been due to s3 having the most proportional bioactive and non-bioactive sequences out

of all the created datasets. The decent accuracy proves that there are dependencies that the LSTM model can see with only the identity of the amino acid and how many of them exist without looking at the position and neighboring amino acids. Figure 4 shows the training graph of s3 while the representative evaluation metrics of s3 are shown in Table 2. Comparing these metrics to the performance of a similar model developed by Tang et al (2022), which also used peptide sequence as an input, their model only achieved an accuracy of 70.9%. Although their output and task are more complex as a multilabel classifier.

Improvements can still be made to the model as it is necessary to increase the amount of non-bioactive peptides to match the number of bioactive peptides. Even if the accuracy of the model is decent without information such as the position of each peptide, it would improve the performance of the model if such information was simultaneously fed through a different type of amino acid encoding method.

4. Conclusion

In this study, an LSTM-based binary classification

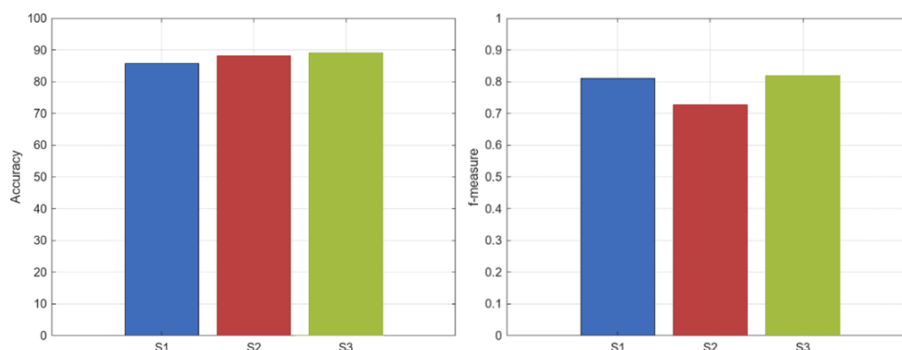


Figure 3. Macro average of the accuracy (L) and the f-measure (R) between models.

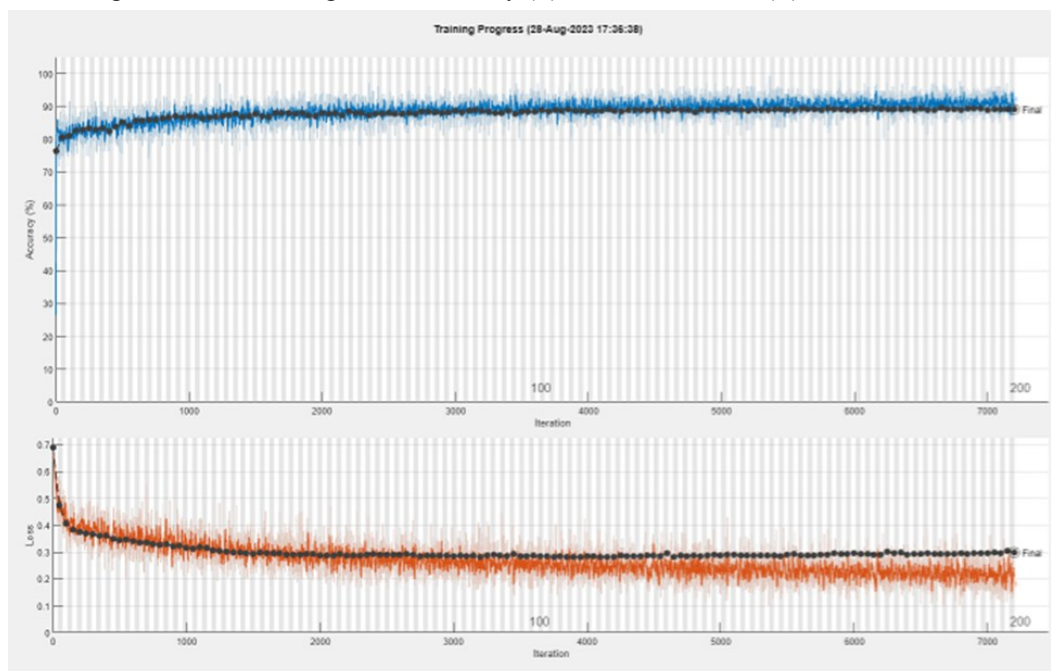


Figure 4. Training graph of the s3 dataset.

Table 2. Evaluation metrics of S3.

S3	Bioactive	Nonbioactive	Macroaverage	Microaverage
true_positive	885	148	516.5	516.5
false_positive	76	50	63	63
false_negative	50	76	63	63
true_negative	148	885	516.5	516.5
precision	0.9209157128	0.7474747475	0.8341952301	0.891285591
sensitivity	0.9465240642	0.6607142857	0.8036191749	0.891285591
specificity	0.6607142857	0.9465240642	0.8036191749	0.891285591
accuracy	89.13%	89.13%	89.13%	89.13%
F-measure	0.9335443038	0.7014218009	0.8174830524	0.891285591

model was built using a combined total of 5795 peptide sequences. The dataset underwent an encoding method similar to the ones used for analyzing documents and was split into 8:2 training and validation datasets, respectively. Two other datasets were also created to compare performance when isolating different lengths of peptides, one where it only consisted of those of length 5 and below and the second for the rest. After training and evaluating the models, the model was trained using the dataset that consisted of all the peptides with an accuracy of 89.13% after 200 epochs. The study demonstrated the possibility of predicting peptide activity just from its sequence. Moreover, this study can be used as a baseline model in creating other frameworks that are capable of predicting peptide activity even in the absence of the original protein sequence, which is a common occurrence in food systems.

Separation of the dataset is not as necessary as initially conceptualized as the model showed similar performance between the three datasets and the best-performing model is the one trained on all the BPs without splitting. It is recommended to build the labeled dataset further by adding more non-bioactive peptide sequences to match the number of bioactive peptides for a more robust model. Giving the model more information, such as amino acid positioning, should also improve the performance of the model. This could be done by using a different type of encoding that preserves more information. Finally, the number of hidden units on the LSTM layer underwent preliminary testing and showed no significant difference between 10, 20, 100, and 200 hidden units; the study opted to stick with 10 hidden units for optimization and resource efficiency as it can make predictions faster. Although fine-tuning the number of hidden units can be used to increase model performance by a few points. The improved model can be used to predict the activity of a peptide even before it is tested experimentally.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

The authors would like to acknowledge the assistance of Mr. Benz Isaac Ong in collecting the data used in this study.

References

- Daliri, E.B., Oh, D.H. and Lee, B.H. (2017). Bioactive peptides. *Foods*, 6, 32. <https://doi.org/10.3390/foods6050032>
- Guo, Y., Cao, X., Liu, B. and Peng, K. (2020). El Niño index prediction using deep learning with ensemble empirical mode decomposition. *Symmetry*, 12, 893. <https://doi.org/10.3390/sym12060893>
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Li, Y., Li, X., Liu, Y., Yao, Y. and Huang, G. (2022). MPMABP: A CNN and Bi-LSTM-based method for predicting multi-activities of bioactive peptides. *Pharmaceuticals*, 15(6), 707. <https://doi.org/10.3390/ph15060707>
- Ma, P., Zhang, Z., Jia, X., Peng, X., Zhang, Z., Tarwa, K., Wei, C.I., Liu, F. and Wang, Q. (2022). Neural network in food analytics. *Critical Reviews in Food Science and Nutrition*, 64(13), 4059-4077. <https://doi.org/10.1080/10408398.2022.2139217>
- Minkiewicz, P., Iwaniak, A. and Darewicz, M. (2019). BIOPEP-UWM database of bioactive peptides: Current opportunities. *International Journal of Molecular Sciences*, 20, 5978. <https://doi.org/10.3390/ijms20235978>
- Mooney, C., Haslam, N.J., Holton, T.A., Pollastri, G. and Shields, D.C. (2013). PeptideLocator: Prediction of bioactive peptides in protein sequences. *Bioinformatics*, 29(9), 1120-1126. <https://doi.org/10.1093/bioinformatics/btt103>
- Naga, I.E. and Murphy, M.J. (2015). What is machine learning? In Naga, I.E., Li, R. and Murphy, M.J. (Eds.) *Machine Learning in Radiation Oncology*, p. 3-11. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-18111-1_1

doi.org/10.1007/978-3-319-18305-3_1

- Sanchez, A. and Vasquez, A. (2017). Bioactive peptides: A review. *Food Quality and Safety*, 1, 29-46. <https://doi.org/10.1093/fqsafe/fyx006>
- Schuster, M. and Paliwal, K.K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681. <https://doi.org/10.1109/78.650093>
- Shi, H. and Zhang, S. (2022). Accurate prediction of anti-hypertensive peptides based on convolutional neural network and gated recurrent unit. *Interdisciplinary Sciences: Computational Life Sciences*, 14, 879-894. <https://doi.org/10.1007/s12539-022-00521-3>
- Tang, W., Dai, R., Yan, W., Zhang, W., Bin, Y., Xia, E. and Xia, J. (2022). Identifying multi-functional bioactive peptide functions using multi-label deep learning. *Briefings in Bioinformatics*, 23(1), bbab414. <https://doi.org/10.1093/bib/bbab414>
- Toldra, F., Reig, M., Aristoy, M. and Mora, L. (2018). Generation of bioactive peptides during processing. *Food Chemistry*, 267, 395-404. <https://doi.org/10.1016/j.foodchem.2017.06.119>
- Udenigwe, C.C. and Aluko, R.E. (2012). Food protein-derived bioactive peptides: Production, processing, and potential health benefits. *Journal of Food Science*, 71(1), R11-R24. <https://doi.org/10.1111/j.1750-3841.2011.02455.x>
- Vidal-Limon, A., Aguilar-Toala, J.E. and Liceaga, A.M. (2022). Integration of molecular docking analysis and molecular dynamics simulations for studying food proteins and bioactive peptides. *Journal of Agricultural and Food Chemistry*, 70(4), 934-943. <https://doi.org/10.1021/acs.jafc.1c06110>