№ 4 (2), 2020 ISSN: 2682 – 9517 (print) ISSN: 2683 – 0930 (online)
INSTITUTE OF INFORMATICS AND INNOVATIVE TECHNOLOGIES

Deep learning algorithm for food-derived antioxidative peptides bioactivity prediction

Margarita Terziyska[1], Ivelina Desseva[2], Zhelyazko Terziyski[3]

[1]Informatics and Statistics Department, UFT Plovdiv, Bulgaria, mterziyska@uft-plovdiv.bg

[2] Department of Analytical Chemistry and Physical Chemistry, UFT Plovdiv, Bulgaria, ivelina_hristova_vn@abv.bg

[3]Faculty of Mathematics and Informatics, University of Plovdiv Paisii Hilendarski, Bulgaria, terziyski@engineer.co

***Abstract*— In the current study it was developed deep learning (DL) architecture - Long Short Term Memory (LSTM) was used for prediction of food-derived antioxidative peptides. The obtained results suggest that the proposed model could be an effective and promising high-throughput method for antioxidant peptides identification.**

Keywords—deep learning, LSTM, antioxidative peptides, prediction.

I. INTRODUCTION

Antioxidants are a group of biochemicals that help the processes in the human body to fight free radicals - atoms or molecules with an unpaired electron that are formed as a side effect of oxidative processes in the body. Antioxidants are compounds that bind to free radicals to neutralize their action and thus reduce factors that could lead to more serious cell damage or disease. Thus, they may prevent or delay some types of cell damage. Free radicals, which lack an electron in the shell, seek to obtain that electron by attacking DNA, fat, and proteins, which in turn are left without an electron, begin to enter into chemical bonds that are harmful to health. Tens of thousands of free radicals are released in our body every minute. The human body produces antioxidants (such as coenzyme Q10, etc.), but with age, deteriorating quality of life or increasing pollution in the environment, the amount and effectiveness of antioxidants produced by our body decreases. then our preventive activity and responsible attitude to the needs of our body becomes of paramount importance. And the latter is expressed in the consumption of rich in antioxidants food within a balanced diet, strengthening the body and supporting its functions by providing natural antioxidants from natural sources - herbs, tea, fruits and more.

Biologically active peptides, in particular those with antioxidant activity, have been successfully used in the fight against free radicals [1–2]. Antioxidative peptides (APs) contain 5–16 amino acid residues [3]. Antioxidant peptides from foods are considered to be safe and healthy compounds with low molecular weight, low cost, high activity, easy absorption [4]. The existing *in vivo* and *in vitro* processes of identifying each individual AP is time-consuming and expensive. Therefore, approaches based on computational methodologies appear to be extremely suitable for predicting APs. In the scientific literature, numerous developments for identification peptides with different activity using machine learning (ML) techniques have been reported. Concrete, for these with antioxidative activity, not much has been done in terms of their prediction with ML techniques.

An effective model for predicting the antioxidant proteins using statistical moments and feed-forward neural networks with gradient descent with adaptive learning rate back- propagation as training algorithm was presented in [5]. A Naive Bayes-based method to predict antioxidant proteins using amino acid compositions and dipeptide compositions was proposed in [6].

Multivariate analyses such as principal component analysis (PCA) was used to cluster the possible amino acid compositions of antioxidant peptides in potato protein hydrolysate (PPH) in [7]. In [8] was presented the AnOxPePred - web-server tool that uses deep convolutional neural network (CNN) to predict the antioxidant properties of peptides.

Since antioxidant activity of peptides is related to the composition and sequence of the amino acids, in this work it was developed deep learning (DL) architecture - Long Short Term Memory (LSTM) as a most appropriate to process entire sequences of data. The gathered from public databases positive and negative peptide sequences are transformed by the Chou's pseudo amino acid composition (PseAAC) method. The obtained features are fed as an input to train the LSTM model. To evaluate the prediction performance of the proposed model, a set of usually applied metrics was used. The obtained results suggest that the proposed method could be an effective and promising high-throughput method for antioxidant peptides identification.

II. BIOACTIVE PEPTIDES PREDICTION PROCESS

Peptides possess specific biological activity, which depends on their structure i.e. on their amino acid sequence. Therefore, the most widely used approach for predicting the BAPs is one based on their sequence. The bioactive peptides (BAPs) prediction process has several stages, which are schematically presented in Fig.1.

A. Dataset preparation

The first stage is related to data preparation. The data can be collected by their function, source, length etc. from public databases [10]. In fact, two datasets are formed - positive and negative. The positive dataset includes known peptide sequences with antioxidant activity. It is more difficult to generate a negative dataset. According to [11], the negative samples should generally consist of the following peptides:

- random peptides retrieved from the UniProt;
- random shuffling of positive samples;
- peptides with different functions rather than the desired function.

The data set are then divided into subsets for training and testing. The proportions are determined depending on the size and type of data available, as well as the model. When using artificial intelligence (AI) model, the most commonly used rule is 80% / 20%, i.e. 80% of the data is used for model training and 20% for evaluation.

In this study, for the construction of the positive antioxidative peptide dataset, the FeptideDB [10] database was used. It was generated 529 positive samples. The negative data set was constructed from peptides without antioxidative function. It consists 734 samples.

B. Feature extraction

In the second stage, a method of encoding the peptides must be selected to obtain digital vectors of a certain length, called descriptors. Thus, encoded data can be used by ML algorithms. In this study, PseAAC, which is a modification on the classic AAC, was adopted as the feature encoding method of peptides [12]. More concrete, the type II PseAAC was applied. In this case, each peptide is represented by a vector with 20+iλ dimensions, where i denotes the number of properties of the amino acid taken into consideration and λ is a coefficient that determines the distance of the interacted amino acids. To generate peptides

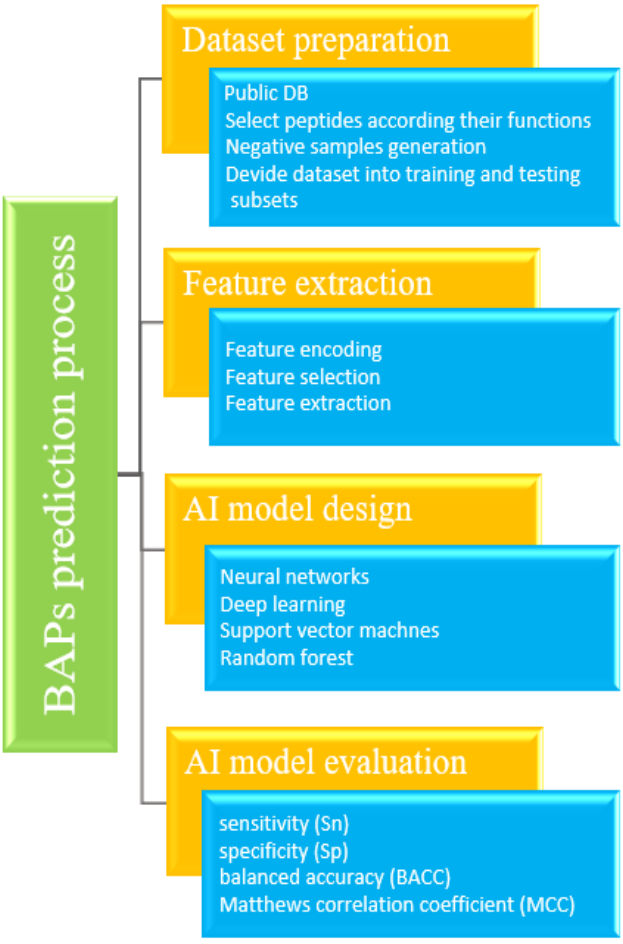

Fig. 1. Predictive AI-based model developing procedure

representation, it was used a web-based application PseAAC [13]. It was chosen the following properties of amino acids: hydrophobicity, hydrophilicity, mass, pK_1 (alpha-COOH), pK_2 (NH_3) and pI (at 25°C). The coefficient λ was set to 1 and the weight factor ω was set to 0.05.

C. AI model design

During the third stage of the process, the predictive model is created. If it is chosen to work with artificial intelligence techniques as in this paper, then the model is constructed based on some of the following algorithms - ANN, DL, SVM, k-NN, RF and etc. The model is trained to make predictions using the prepared set (positive and negative) of data together with the descriptors of physicochemical characteristics.

D. AI model evaluation

The process of predicting the biological activity of peptides ends with the evaluation of the model. In the present study, four performance indicators are used - accuracy (Acc), sensitivity (Sn), specificity (Sp) and Matthew correlation coefficient (MCC) [14,15]. They are calculated with the following expressions:

$$Acc = \frac{TN + TP}{TN + TP + FN + FP} \quad (1)$$

$$Sn = \frac{TP}{TP + FN} \quad (2)$$

$$Sp = \frac{TN}{TN + FP} \quad (3)$$

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \quad (4)$$

where TP is the number of positive samples, TN is the number of negative samples, and FP and FN are the misclassified positive and negative samples, respectively.

III. RESULTS

The structure of the proposed LSTM model consists a LSTM layer with 64 memory cells (see fig. 2). Additionally, to LSTM layer there is a dropout layer and a dense layer. A Dropout layer is used as a regularization method where input and recurrent connections to LSTM units are probabilistically excluded from activation and weight updates while training a network. This has the effect of reducing overfitting and improving model performance. A Dense layer feeds all outputs from the previous layer to all its neurons, each neuron providing one output to the next layer.

The LSTM model was built with the Keras framework (www.keras.io) using a sequential model and a TensorFlow [16] deep learning library back-end. The Keras model was compiled using 50 epochs with the 'adam' optimizer (loss: binary_crossentropy, metrics: accuracy). The model accuracy and loss are shown on fig. 2 and fig. 3, respectively.

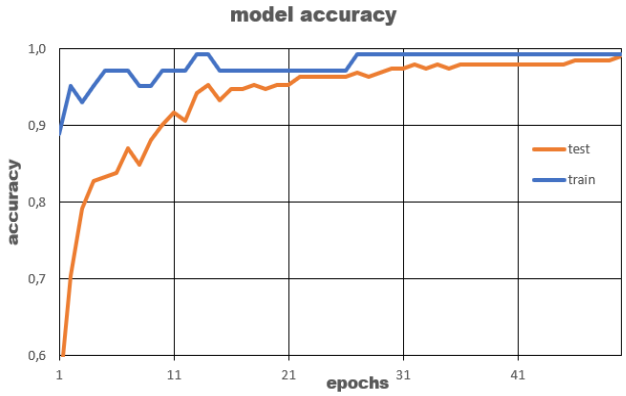

Fig. 2. Model accuracy

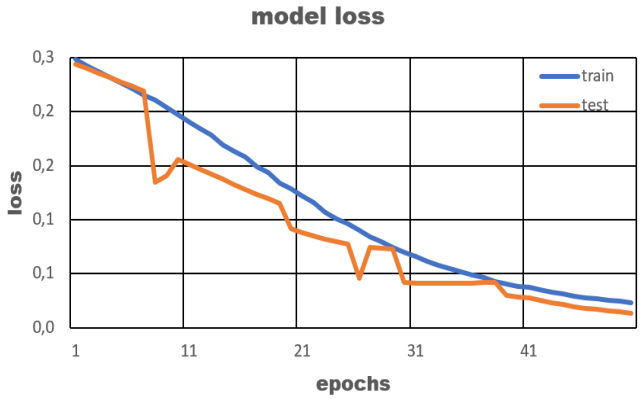

Fig. 3. Model loss

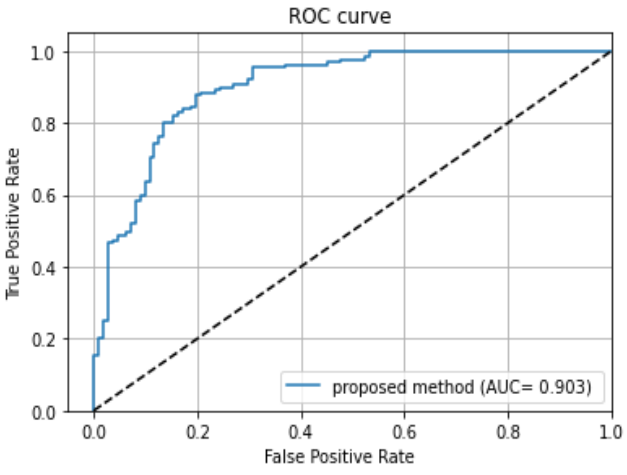

Fig. 4. The performance of the proposed LSTM model

The area under the receiver operating characteristic (ROC) curve (AUC) also is calculated to evaluate the performance of models. It is shown on fig. 4. The range of AUC value is from 0 to 1, and a perfect classifier can be found as AUC=1.0 while the classifier has no discriminative power as AUC=0.5. In this study, the AUC is 0.903, which is close to the perfect classifier.

The 5-fold cross-validation details are presented in Table 1. The average accuracy of 5-fold cross-validation was 83.33%, the average sensitivity (Sen) was 89.17%, the average specificity (Spec) was 76.64%, the mean precision (Prec) was 82.14% and the Matthews correlation coefficient (MCC) was 67.37%

TABLE I. PERFORMANCE OF LSTM MODEL DURING THE 5-FOLD CROSS-VALIDATION

Fold set	Acc(%)	Prec(%)	Sen(%)	Spec(%)	Mcc(%)
1	91.65	86.64	96.01	81.84	84.21
2	79.17	75.00	92.31	63.62	59.13
3	81.25	81.48	84.61	77.27	62.16
4	83.33	90.91	76.92	90.91	67.83
5	81.25	76.67	92.00	69.56	63.52
Average	**83.3±4.9**	**82.1±6.7**	**89.2±7.6**	**76.64±10.6**	**67.37±9.9**

IV. CONCLUSIONS

Numerous scientific studies show that there is a direct relationship between oxidative stress and a number of diseases, such as type 2 diabetes, cardiovascular disease, cancer, etc. Oxidative stress can be controlled with the help of antioxidants isolated and identified from natural sources (e.g., food). Several methods have been developed to evaluate the antioxidant efficacy of foods, but they mainly use *in vitro* and *in vivo* protocols. The latter are too expensive and time consuming. Therefore, in recent years the aim is to develop computer computational methods for predicting the biological activity of peptides. In view of this, the present study presents such a method based on deep networks. An LSTM architecture has been developed and used to predict the antioxidant activity of peptides. The analysis shows that the deep learning algorithm has high predictive accuracy, which make it suitable for identification of antioxidative peptides.

ACKNOWLEDGMENT

The authors acknowledge the support received from the Bulgarian Scientific Fund – contract № KP-06-M36/2.

REFERENCES

[1] Sohaib, M., Anjum, F.M.; Sahar, A.; Arshad, M.S.; Rahman, U.U.; Imran, A.; Hussain, S. Antioxidant proteins and peptides to enhance the oxidative stability of meat and meat products: A comprehensive review. Int. J. Food Prop. 2017, 20, 2581–2593.

[2] Lorenzo, J.M., Munekata, P.E.S.; Gómez, B.; Barba, F.J.; Mora, L.; Pérez-Santaescolástica, C.; Toldrá, F. Bioactive peptides as natural antioxidants in food products – A review. Trends Food Sci. Technol. 2018, 79, 136–147, doi:10.1016/j.tifs.2018.07.003.

[3] Chen H M, Muramoto K, Yamauchi F, Nokihara K. "Antioxidant activity of design peptides based on the antioxidative peptide isolated from digests of a soybean protein" in J Agric Food Chem 1996;44:2619–23.

[4] Sarmadi, B. H., & Ismail, A. "Antioxidative peptides from food proteins: a review", in Peptides, 2010, 31(10), 1949-1956..

[5] Butt, A. H., Rasool, N., & Khan, Y. D. "Prediction of antioxidant proteins by incorporating statistical moments based features into Chou's PseAAC", in Journal of theoretical biology, 2019, 473, 1-8.

[6] Feng, P. M., Lin, H., & Chen, W. ""dentification of antioxidants from sequence information using naive Bayes", In Computational and mathematical methods in medicine, 2013.

[7] Wu, J., Mao, C., Zhang, W., & Cheng, Y. "Prediction and Identification of Antioxidant Peptides in Potato Protein Hydrolysate", Journal of Food Quality, 2020.

[8] Olsen, T. H., Yesiltas, B., Marin, F. I., Pertseva, M., García-Moreno, P. J., Gregersen, S., ... & Marcatili, P. "AnOxPePred: using deep learning for the prediction of antioxidative properties of peptides". In Scientific Reports, 2020, 10(1), 1-10.

[9] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[10] Panyayai, T., Ngamphiw, C., Tongsima, S., Mhuantong, W., Limsripraphan, W., Choowongkomon, K., Sawatdichaikul, O.: FeptideDB: A web application for new bioactive peptides from food protein. Heliyon, 5(7), (2019).

[11] Basith, S., Manavalan, B., Hwan Shin, T., Lee, G.: Machine intelligence in peptide therapeutics: A next - generation tool for rapid disease screening. Medicinal research reviews (2020).

[12] Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo - amino acid composition. Proteins: Structure, Function, and Bioinformatics, 43(3), 246-255.

[13] Shen, H. B., & Chou, K. C.: PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Analytical biochemistry, 373(2), 386-388 (2008).

[14] Zhang, L., Zhang, C., Gao, R., Yang, R., Song, Q. (2016). Sequence based prediction of antioxidant proteins using a classifier selection strategy. Plos one, 11(9), e0163274.

[15] Damyanov C., Analysis and Evaluation of Errors in Training and Testing of Recognition Systems. ICCST, 1(1), pp. 13-19, 2012

[16] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16) (pp. 265-283).