Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 198 (2022) 480–485

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Internet of Smart Things (IST 2021)
November 1-4, 2021, Leuven, Belgium

Protein Post-translational Modification Site Prediction using Deep Learning

Yujuan Deng[a,b], Yunfang Fu[a,b], Huitao Zhang[a,b,*], Xin Liu[c], Zhiguo Liu[a,b]

[a]*School of Computer Science and Engineering, Shijiazhuang University, Shijiazhuang 050035, China*
[b]*Hebei Province Internet of Things Intelligent Perception and Application Technology Innovation Center, Shijiazhuang 050035, China*
[c]*Shijiazhuang Public Security Bureau, Shijiazhuang 050035, China*

Abstract

The study and identification of protein post-translational modification(PTM) sites in rice seeds plays an important role in breeding and yield increasing in intelligent agriculture. Deep learning, especially deep convolutional neural networks(CNN) become more and more popular in intelligent agriculture because of its excellent image processing and data analysis capabilities. In this paper, the Denoised-Oversampling is adopted to denoise the samples and balance the data. The protein sequences are transformed into corresponding numerical feature vectors. Then, the processed feature subset is input to the deep convolution network. In fully connected layers, we used the L2 regularization method to prevent the overfitting problem. The phosphorylation sites are predicted by the sigmoid classifier. Finally, in the experimental part, in order to truly reflect the predictive power of the model, we divided the dataset into training set, validation set and independent test set. The prediction accuracy of the independent test set is 83.20%. The test result shows that the proposed algorithm has better prediction performance than the existing phosphorylation prediction algorithms.

Keywords: Deep Learning；Convolutional Neural Networks；Intelligent Agriculture；Modification Site Prediction；

* Corresponding author.
E-mail address: 12116317@bjtu.edu.cn

10.1016/j.procs.2021.12.273

1. Introduction

Rice is an important staple food supporting more than half of people all over the world. High yield is one of the eternal goals of breeders and farmers. Seed germination has been reported to affect vigor and crop performance, which in turn influences yield. Acetylation, ubiquitylation, succinylation, phosphorylation, sumoylation, carbonylation and glycosylation have been reported to regulate protein activities in the stage of seed germination[1]. Among them, phosphorylation is a prevalent PTM activity and plays a vital role during seed germination. Phosphorylation site has a great impact on the protein activity. By far, there were only few studies on the PTM during the rice seed germination. It is hard to comprehensively identify the phosphorylated proteins in traditional methods because of the relatively low abundance and readily degradable characteristics of the proteins in seeds. In recent years, machine learning and deep learning have been used in predicting PTM of protein phosphorylation in plants[2]. But protein sample has its own limitations. First of all, the sample size is limited. Secondly, the composition structure of 20 common amino acids determines its own unique coding method. Therefore, how to analyze, study and forecast characteristics accurately by deep learning model is a challenge. CNN can analyze the complex characteristics of the sequence, so as to establish the model to predict the sites of protein sequence, which is becoming a popular research method. Recent years, CNN was used in DeepBind for predicting sequence specificities of DNA- and RNA-binding proteins[3]; Duolin Wang developed the MusiteDeep framework[4-5], which provided the phosphorylation site prediction for humans; Lin proposed a new rice-specific svm predictor for protein phosphorylation sites[6]. In this paper, the method of phosphorylation site prediction for rice is based on CNN.

2. Methods

In this paper, the Denoised-Oversampling is adopted to denoise the samples and balance the data. The protein sequences are transformed into corresponding numerical feature vectors by One-hot Encoding. Then, the processed feature subset is input to the deep convolution network. The specific methods are as follows.

2.1. Data preprocessing method of Denoised-Oversampling

We selected the aba-induced protein phosphorylation in rice. The phosphorylation sites and sequence information were collected to construct the database. In the protein substrate, sequence fragments with phosphorylated annotation sites on serine (S) were taken as positive samples, and sequence fragments without phosphorylated annotation sites were taken as negative samples. In reality, the numbers of positive and negative samples are extremely unbalanced. Only a small percentage of the protein sequence are phosphorylated. Therefore, the classifier may ignore the influence brought by the small number of samples and overemphasize the accuracy of the most samples, which will make the classification a counterexample and lead to the failure of classification. For example, the ratio of positive and negative samples is about 1:45 in our database. Moreover, at the terminal of the protein sequence, the length of the truncated sequence fragment is less than 21, which is generally completed with X in biology. In the training model, the meaningless X adds a lot of noise. To balance the data, Chawla studied the SMOTE to composite minority samples[7]. Han used a new oversampling method called Borderline-SMOTE to balance data sets[8]. To solve the problems above, we carried out the method of Denoised-Oversampling, which consists of two steps: denoising and data balancing. Firstly, in order to avoid the impact of noise on training, we delete the sequences containing X. Secondly, we traverse 45 times for the positive samples to balance the number of positive and negative samples.

2.2. One-hot Encoding

In order to facilitate the input of the model, one-hot encoding[9] is widely used. The categorical features are converted to k numerical features, which means that the categorical features of non-numerical type are quantized into numerical type. By this way, the transformed format cooperates well with the classification algorithm. In the

method, only the value of the place which corresponding to the amino acid is marked as 1 and the others as 0. There are 20 common amino acids, so the value of the k is 20. According to this principle, we get a one-of-20 code as the input. The one-hot encoding is shown as Fig. 1.

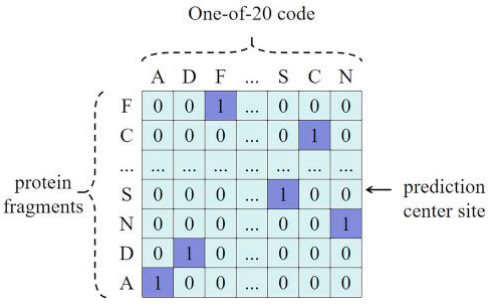

Fig. 1. One-of-20 code.

2.3. Building the CNN

The input matrix processed by one-hot encoding is sent to the CNN for learning through training. The training model employed in this paper consists of two convolutional layers and two fully connected layers. The training model is shown as Fig. 2.

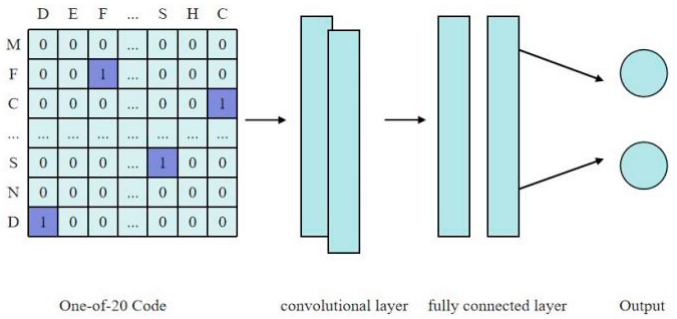

Fig. 2. Deep-learning architecture of our method.

We use convolution layer to extract local features better. The important part of CNN is that convolution operation is used by filtering the input matrix. Compared to AlexNet[10] whose filters have large dimension, the filters with smaller dimension can effectively reduce the parameter magnitude. At the same time, the architecture of small filters can improve the resolution of each class.

There is a ReLU(Rectified Linear Units) layer closely behind the convolution layer[11-12]. We have designed two fully connected layers respectively to classification and regression. In order to alleviate overfitting, L2 regularization is used[13]. It adds the L2 norm of the parameter to the loss function, which can make the parameter smaller when the gradient decreases, so as to prevent the model from fitting a particularly complex function. Meanwhile, we adopted gaussian truncation distribution to initialize the standard deviation[14]. Finally, sigmoid is used to output the binary classification results. The structure and parameters are as follows:

Input layer: a protein fragment encoded by one-of-20.

Layer 1: 1D convolutional layer. It consists of 10 1D convolution kernels whose length is 2 and step size is 1. The activation function is ReLU.

Layer 2: 1D convolutional layer. It consists of 10 1D convolution kernels whose length is 3 and step size is 1. The activation function is ReLU.

Layer 3: Fully connected layer. There are 64 neurons. L2 regularization parameter is 0.01. The mean value of gaussian truncation distribution is 0 and the standard deviation is 0.01. Dropout mechanism random deletion rate is 0.5.

Output layer: Sigmoid output layer. It's made up of two neurons. L2 regularization parameter is 0.01. The mean value of gaussian truncation distribution is 0 and the standard deviation is 0.01.

3. Experiments and Results

In this part, the source and segmentation of the database are introduced in detail. At the same time, we adopt a variety of indicators to evaluate the experiments compared with other classical models. The results show that our model has advantageous.

3.1. Database

In the experiment, we divided the samples into training set, validation set and independent test set to ensure the real effect of the model, as shown in Table 1. Supplementary database can be found at www.mdpi.com/1422-0067/18/1/60/s1.

Table 1. Database.

Category	Original data	X-deleted data	Denoised-Oversampling	Train	Validation	Independent test
Positive	2561	2366	83970	56250	27720	500
Negative	89377	86148	85648	57384	28264	500

3.2. Evaluation indicators

In our study, seven indicators are adopted to evaluate the performance including Recall_score, F1_score, Precision_score, Roc_auc_score, Train-accuracy, Valid-accuracy and Test-accuracy. In particular, the Receiver Operating Characteristic (ROC) curve can consider the accuracy of positive and negative samples at the same time and ignore the imbalance problem of samples. It uses false positive rate as the horizontal axis and true positive rate as the vertical axis. Area Under Curve (AUC) is the area enclosed by the ROC curve and the coordinate axis. The closer the AUC is to 1.0, the better performance the model has. When the AUC is equal to 0.5, the authenticity is the lowest and has no application value. Therefore, we selected the ROC curve to make a more objective evaluation of the experimental results.

3.3. Comparison of different treatment methods for sample imbalance problem

To solve the problem of data imbalance, Random-Undersampling and SMOTE are used to compare with our method named Denoised-Oversampling. The first two methods scored poorly on independent test sets. It shows that the generalization ability of the above methods are not satisfied. Compared with the first two methods, the accuracy of Denoised-Oversampling increased on the training set and validation set. Meanwhile, the performance on the independent test set is also the best. Experimental results are shown in Table 2.

Table 2. Experimental results of different methods to solve the problem of sample imbalance.

Evaluation indicators	Random-Undersampling	SMOTE	Denoised-Oversampling
Recall_score	0.5785	0.8826	**0.9140**
F1_score	0.6606	0.8683	**0.8747**
Precision_score	0.7699	0.8544	0.8387
Roc_auc_score	0.7651	0.9410	0.9372
Train-accuracy	0.7220	0.8491	**0.8632**
Valid-accuracy	0.6948	0.8666	**0.8707**
Test-accuracy	0.6850	0.6000	**0.8320**

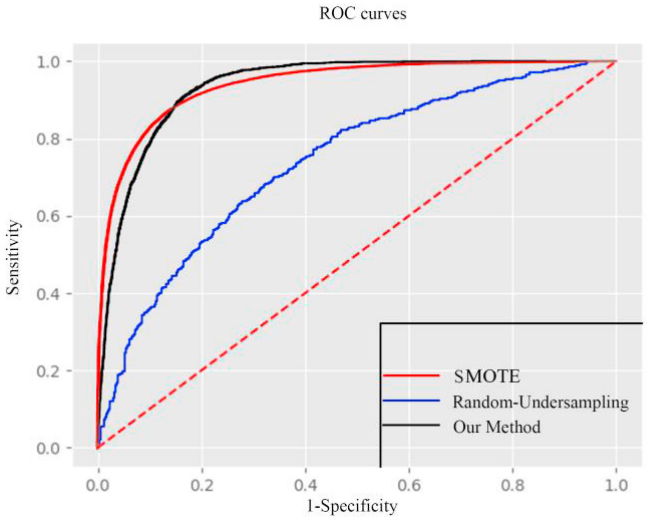

Fig. 3. ROC curves of different methods

ROC curves on the validation set are shown in Fig. 3. Although the ROC curve of SMOTE performed slightly better, the accuracy is not satisfied on the independent test set. This is because that the SMOTE samples are based on distance measure, which inevitably generate noise. Although the classification effect is satisfied on the validation set which means the AUC value is high, the accuracy is poor on the independent test set. The test_accuracy is only 62.2%. Denoised-Oversampling performed best on the independent test set. The test_accuracy is 83.2%, which shows strong generalization ability.

3.4. Compared with other structures

In order to avoid network degradation when more convolutional layers are introduced, Residual Network (ResNet)[15] are widely used. In the building block of the ResNet, the identity mapping is added while the number of network layers is added to the structure. The principle of the ResNet is that when the number of feature maps is halved, the number of convolution kernels is doubled. Therefore, the total number of learnable parameters remains unchanged. By this way, the result of the network is not worse than original. In the experiment, our method is compared with the ResNet-18. The ResNet-18 consists of 17 convolutional layers and one fully connected layer. Although it performed well on the training set and validation set, it did not achieve good results on the independent

test set. It shows that the generalization ability of the ResNet-18 is not satisfied. Compared with ResNet-18, the proposed algorithm in this paper performed better on generalization. Experimental results are shown in Table 3.

Table 3. Experimental results of ResNet-18 and our method.

Evaluation indicators	ResNet-18	Our Method
Recall_score	1.0	0.9140
F1_score	0.9968	0.8747
Precision_score	0.9938	0.8387
Train-accuracy	0.9998	0.8632
Valid-accuracy	0.9972	0.8707
Test-accuracy	0.523	**0.8320**

4. Conclusions & Future Work

In this paper, we constructed a CNN model to predict protein PTM sites in rice. Compared with other existing methods, our model obtained better prediction result on the independent test set. However, how to extract features more accurately and deal with the problem of imbalance samples better are still subjects that need further research.

Acknowledgment. The work is supported by the (211080251a) project.

References

[1]He, D., Wang, Q., Li, M., Damaris, R. N., Yi, X., Cheng, Z., & Yang, P. (2016). Global proteome analyses of lysine acetylation and succinylation reveal the widespread involvement of both modification in metabolism in the embryo of germinating rice seed. Journal of proteome research, 15(3), 879-890.

[2]Yu, F., Li, M., He, D., & Yang, P. (2021). Advances on Post-translational Modifications Involved in Seed Germination. Frontiers in Plant Science, 12, 362.

[3]Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology, 33(8), 831-838.

[4]Wang, D., & Liu, D. (2017, November). MusiteDeep: A deep-learning framework for protein post-translational modification site prediction. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2327-2327). IEEE.

[5]Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., & Xu, D. (2017). MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics, 33(24), 3909-3916.

[6] Lin, S., Song, Q., Tao, H., Wang, W., Wan, W., Huang, J., ... & He, H. (2015). Rice_Phospho 1.0: A new rice-specific SVM predictor for protein phosphorylation sites. Scientific reports, 5(1), 1-9.

[7]Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[8]Han, H., Wang, W. Y., & Mao, B. H. (2005, August). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing (pp. 878-887). Springer, Berlin, Heidelberg.

[9] Alkharusi, H. (2012). Categorical variables in regression analysis: A comparison of dummy and effect coding. International Journal of Education, 4(2), 202.

[10]Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[11]Nair, V., & Hinton, G. E. (2010, January). Rectified linear units improve restricted boltzmann machines. In Icml.

[12]Li, Y., & Yuan, Y. (2017). Convergence analysis of two-layer neural networks with relu activation. arXiv preprint arXiv:1705.09886.

[13]Cortes, C., Mohri, M., & Rostamizadeh, A. (2012). L2 regularization for learning kernels. arXiv preprint arXiv:1205.2653.

[14]Burkardt, J. (2014). The truncated normal distribution. Department of Scientific Computing Website, Florida State University, 1-35.

[15]He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).