Prediction of LEA Plant Proteins Based on Machine Learning

Siquan Hu1,2*, Meidi Zhao2, Zhiguo Shi2 and Min Zhang3

1 Shunde Graduate School, University of Science and Technology Beijing, Foshan, China
2 School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
3 School of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu, China
Email: husiquan@ustb.edu.cn

Abstract. Plants’ ability to grow smoothly in harsh environments is restricted by various factors. Late embryogenesis abundant proteins play a vital role in plants’ resistance to environmental stresses. Therefore, identifying lea proteins has become a valuable task. Although biological experiments are very important in such task, machine learning methods could narrow the searching scale with a certain accuracy and the increased speed. At present, there is a lack of such work in this area. This paper proposes a set of machine learning methods for predicting late embryogenesis abundant proteins and compares the performance of SVM and random forest (RF) methods with CNN method. Samples are collected from UniProt and NCBI with experts’ confirmation. The results show that the RF model is better than other models. The accuracy of the RF model is higher than 89% on the training set and is higher than 91% on different independent samples. The work provides some new ideas and methods for the identification of LEA proteins.

1. Introduction
Drought, salinization and freezing are the main stress factors restricting plant growth and development. For instance, drought is the main abiotic stress in cotton production, which can severely hinder plant growth and lead to a decline in yield. Although cotton is well tolerated by water scarcity when the water supply is restricted, the optimal growth and yield of cotton may be adversely affected [1]. Late embryogenesis abundant (LEA) protein is an important plant cell dehydration protective protein, which exerts an enormous function on resisting outside pressure such as drought [2]. The LEA protein was originally found in cottonseed and later in many other species. At present, the research on LEA protein has been carried out for more than 20 years and the consensus has been reached that LEA protein is closely related to abiotic stress tolerance, especially dehydration and cold stress. However, LEA proteins’ structure and functional mechanisms have not been fully understood, which has led to the protein family being called “unsolved mysteries” or “mysteries”.

To enhance the understanding of LEA proteins’ functional mechanisms, a nature first step will be to identify LEA proteins from species as many as possible and find out their similarity. Currently, most LEA proteins identification is carried out by traditional experimental methods, namely a mass spectrometry method based on the MS MOWSE, MASCOT, ProFound16-4, and so on. Their disadvantage is that the process is slow and expensive with complex steps. Thus, identifying plants’ lea proteins quickly is a challenge in the research of LEA protein family and has an important value for
future solving of the mysteries.

In recent years, some protein studies have applied machine learning methods and achieved promising results. In 2001 Hua [3] studied protein sub cells by the SVM method and obtained a good prediction success rate. In 2011 Muppirala et al. [4] proposed the RPISeq (RNA-protein interaction using sequence) method to extract the sequence characteristics of proteins and RNA, and used support vector machine (SVM) algorithm and random forest (RF) algorithm to identify NCRPI of non-coding RNA-protein interaction respectively. In 2016, Li et al. [5] applied deep learning to the study of the secondary structure of proteins and obtained good research results.

Since the function of a protein is mainly determined by its amino acid sequence, computational analysis of confirmed LEA proteins' amino acid sequence could reveal the common feature and help on the prediction of unconfirmed LEA proteins. This paper applies machine learning methods on the prediction of LEA proteins to improve the prediction efficiency. The SVM, CNN, and RF models are built for LEA protein prediction, the models are trained on the training set and the accuracy rates are compared on the test set.

2. Materials and Methods

2.1. Data

UniProt is a comprehensive resource of protein sequence and annotation data [6]. NCBI (https://www.ncbi.nlm.nih.gov/) is also a rich library of protein sequences. We searched by keywords from these two databases and collected relevant data. These data are highly reliable proteins screened by experts.

As shown in table 1, the data set used in the plant LEA proteins model has 90 positive samples. Due to the small amount of LEA protein in the plant kingdom, the proportion of negative data sets has been expanded and 240 negative samples labeled by experts as other kinds of proteins are used in data selection. Both unbalanced datasets and a balanced data set are used, where in either case 70 positive samples were used as the train set, and 20 positive samples were used as the test set. In the train set of the unbalanced dataset, the amount ratio of positive samples to negative samples is 1:2. In both datasets, another test set2 is setup by expanding the negative samples' number to evaluate the models’ performance.

Table 1. Datasets for LEA protein prediction models.

| Dataset          | Positive samples | Negative samples | Total |
|------------------|------------------|------------------|-------|
| Original set     | 90               | 240              | 330   |
| Balanced dataset |                  |                  |       |
| Train set        | 70               | 70               | 140   |
| Test set         | 20               | 20               | 40    |
| Test set_2       | 20               | 100              | 120   |
| Unbalanced dataset |                |                  |       |
| Train set        | 70               | 140              | 210   |
| Test set         | 20               | 40               | 60    |
| Test set_2       | 20               | 100              | 120   |

2.2. Methods

2.2.1. Different Machine Learning Models. Machine learning is the general term for algorithms that attempt to extract hidden patterns from large amounts of historical data and use them to predict or classify them. The support vector machine (SVM) is a learning method carried out by Vapnik [7] based on statistical theory to restrict training errors and minimize confidence range in the case of limited samples. The random forest algorithm (RF) was first proposed by Breiman [8] and it is an integration of
decision trees whose construction is determined by a random vector. Logistic regression (LR) is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome [9].

According to previous studies, it is clear that the classification of a protein has an obvious relationship with its amino acid composition. The pseudo-amino acid composition method proposed by Chou is to express the protein sequence with a 20+ lambda vector [10], where the λ dimension is the frequency at which 20 amino acids are present in the sequence. Formula 1 gives the set of abbreviated capital letters of 20 amino acids. Because of the importance of hydrophilicity and hydrophobicity in LEA proteins, the conservation of hydrophilic and hydrophobic distribution in proteins can be used to describe the characteristics of protein.

\[
\text{AA} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}
\] (1)

According to hydrophilicity, 20 kinds of amino acids were reduced to 6 types [11]. Strong hydrophilic or polar classes, including R, D, E, F, G, and H, are expressed by Q; Strong hydrophobic classes, including L, I, V, A, M, and F are represented by S; Weakly hydrophilic or weakly hydrophobic, including S, T, Y, and W are represented by A; three special amino acids: Proline, Glycine and Cysteine are represented by P, G, and C respectively. In this way, a total of 26-dimension feature vector in Formula 2 is used to extract protein features, where \( j \) from 1 to 26 represents 20 different amino acids and 6 reduced amino acid categories. The \( \zeta \) is deduced according to the protein category, LEA proteins and non-LEA proteins. The factor \( k \) is the sequence number of class \( \zeta \) protein, \( T \) means matrix transpose, \( x_{j,k}^\zeta \) means the frequency of occurrence of the \( j \)th amino acid composition of the \( k \)th sequence of type \( \zeta \). \( m \) represents the total number of sequences in different categories.

\[
x_k^\zeta = [x_{1,1}^\zeta, x_{1,2}^\zeta, \ldots, x_{1,j}^\zeta, \ldots, x_{20,1}^\zeta, x_{20,2}^\zeta, \ldots, x_{20,26}^\zeta]^T, (j = 1, 2, \ldots, 26; \zeta = 1, 2; k = 1, 2, \ldots, m)
\] (2)

2.2.2. CNN Model. The CNN model is composed of a coding layer, an embedding layer, three convolution layers, a flatten layer and a dense layer as shown in figure 1. The role of the coding layer is to use specific numbers to represent each different amino acid. For example, [M, S, A] is coded as [13, 16, 1]. Afterward, it is vectorized by the embedding layer to facilitate feature extraction. Then there are three convolution layers, each of which uses 128 filters and is followed by a pooling layer with 2 pooling unit parameters. The last output layer uses the sigmoid function.

![Figure 1. Structure of the CNN model.](image)

2.2.3 Cross-Validation. In this paper, we carried out cross experiments both on balanced dataset and unbalanced dataset. The parameter \( k \) and 1-\( k \), represent the percentage of training data and verification data in the total model data respectively. Based on experience, \( k \) is set as 0.8, 0.85, 0.9. The results after many trials show that the model can fit the data better and predict the sequence accurately when \( k = \)}
For the parameters of the SVM model and the RF model, we made several choices as shown in table 2, and conducted experiments separately. The choice of the optimizer is also related to the quality of the CNN model, so we chose two optimizers, Adam and SGD. Therefore, we made a total of 60 experimental areas to compare the models with different parameters.

| Experimental group | SVM- parameters | RF- parameters | CNN- optimizers |
|--------------------|----------------|----------------|----------------|
| Balanced data      | k = 0.8, 0.85, 0.9 | rbf, sigmoid, linear, poly | Adam, SGD |
| Unbalanced data    |                | 100, 10, 20, 30 | |

### 3. Result

Based on experiments, the parameters with the best effect are determined in the final models. The performance results are showed in table 3. Both the accuracy on the training set and the accuracy on the test set are better in the unbalanced dataset than those in the balanced dataset respectively. So the following conclusions are mainly based on the unbalanced dataset. Comparing the four models, we can see that the test accuracy of the SVM model is similar as the LR model, while the CNN model is slightly lower, and the RF model is 3.3% higher than the SVM model. So, the RF model has the best accuracy of LEA protein prediction, reaching 91.6%.

| Experimental category | Model | Train-Acc | Test-Acc |
|-----------------------|-------|-----------|----------|
| Balanced experiment   | SVM   | 79.1%     | 72.5%    |
|                       | LR    | 76.4%     | 77.5%    |
|                       | RF    | 85.7%     | 85.0%    |
|                       | CNN   | 93.75%    | 85.0%    |
| Unbalanced experiment | SVM   | 84.3%     | 88.3%    |
|                       | LR    | 84.8%     | 86.6%    |
|                       | RF    | 89.8%     | 91.6%    |
|                       | CNN   | 90.6%     | 68.3%    |

### 4. Discussion

Since LEA protein family is a tiny part of the whole protein world, we did a separate test on Test set_2 in which negative samples are expanded. The results are illustrated in figure 2, and more indicators are examined such as the overall predictive accuracy (ACC), Recall (Rc), Precision (Pr), F1-score, and Matthews correlation coefficient (MCC) [12-13]. The RF model has the best performance on all the five indicators. The SVM model is similar to the LR model, and the CNN model is similar to the RF model on Rc. Therefore, we believe that the RF model is the best predictors among the four examined models. And the RF model has a good performance in the prediction of Non-LEA protein and the accuracy rate can reach 95%, which is similar to the results of the SVM and the LR model. The accuracy of the CNN model on this issue is 80%, which is lower than other models. It can be explained as the current confirmed LEA protein number is very limited and the deep learning potential has not been fully utilized.
5. Conclusion
In this work, we setup four machine learning models for LEA protein prediction problem including SVM, LR, RF and CNN. The experts' confirmed LEA proteins are used as positive samples in the model training. Balanced and unbalanced datasets with different configuration and indicators are used in the performance comparisons. By combining the hydrophobic characteristics of LEA protein to extract 26 features for use, the RF model is superior to the SVM, LR, and CNN models. Thus, we propose a method for predicting LEA protein based on the RF model to easy the future LEA protein research.

Acknowledgments
This work is supported by the National Natural Science Foundation of China (Grant No. 61977005), the Sichuan Science and Technology Program (No. 2018GZDZX0034) and the Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB (BK19CF003).

References
[1] Kaiser W M 2006 Effects of water deficit on photosynthetic capacity Physiologia Plantarum 71 142-149.
[2] Saucedo A L, Hernández-Domínguez E E, de Luna-Valdez L A, Guevara-García A A, Escobedo-Moratilla A, Bojorquéz-Velázquez E, Del Río-Portilla F, Fernández-Velasco D A and Barba de la Rosa A P 2017 Insights on structure and function of a late embryogenesis abundant protein from Amaranthus cruentus: An intrinsically disordered protein involved in protection against desiccation, oxidant conditions, and osmotic stress Frontiers in Plant Science 8 497.
[3] Hua S J and Sun Z R 2001 Support Vector machine approach for protein subcellular location of prediction Bioinformatics 17 721-728.
[4] Muppirala U K, Honavar V G and Dobbs D 2011 Predicting RNA-protein interactions using only sequence information BMC Bioinformatics 12 489.
[5] Li Z and Yu Y 2016 Protein secondary structure prediction using cascaded convolutional and recurrent neural networks International Joint Conference on Artificial Intelligence (IJCAI).
[6] Michele M and Uniprot C 2011 Uniprot knowledgebase: a hub of integrated protein data Database.
[7] Vapnik V N 2000 The Nature of Statistical Learning Theory pp 15-52.
[8] Breiman L 2001 Random forests Machine Learning 45 5-27.
[9] Hosmer D W and Lemeshow S 2000 Applied Logistic Regression pp 25-45.
[10] Chou K C and Elrod D W 1999 Protein subcellular location prediction Protein Engineering 2 2.
[11] Panek J, Eidhammer I and Asland R 2005 A new method for identification of protein families in a set of proteins based on hydropathy distribution in proteins Proteins 58 923-934.
[12] Matthews B W 1975 Comparison of the predicted and observed secondary structure of T4 phage lysozyme Biochimica et Biophysica Acta 405 442-451.
[13] Ilham A S, Ramzan K U and Victor V S 2017 Tssplant: a new tool for prediction of plant pol II promoters Nucleic Acids Research 45 65.