Eco. Env. & Cons. 28 (August Suppl. Issue) : 2022; pp. (S493-S496)
Copyright@ EM International
ISSN 0971–765X

DOI No.: http://doi.org/10.53550/EEC.2022.v28i04s.075

Classification of LEA proteins using Support Vector Machine Algorithm

S. Mahalakshmi*[1], R. Pangayar Selvi[2], V. Anandhi[3] and N. Bharathi[4]

[1,2,3] Department of Physical Sciences & Information Technology, AEC & RI, Tamil Nadu Agricultural University, Coimbatore 641 003, T.N., India
[4]Department of Plant Molecular Biology and Bioinformatics, AC & RI, Tamil Nadu Agricultural University, Coimbatore 641 003, T.N., India

(Received 18 February, 2022; Accepted 22 June, 2022)

ABSTRACT

The tolerance of plants to environmental stress is aided by numerous proteins, one such protein is a Late Embryogenesis Abundant (LEA) Protein. Thus, it is very essential to identify and classify LEA proteins. For this purpose, machine learning would be greatly useful. In this study, an attempt was made to develop SVM classifiers for the classification of LEA proteins based on their physico-chemical properties. The physico-chemical properties were analyzed from protein sequences and then the dataset is split into training and test data. The model is developed and evaluated using cross-validation techniques like 5-fold. The results show that the accuracy of developed SVM classifiers ranges from 60 to 97 per cent.

Key words : *Support Vector Machine, LEA proteins, Cross-validation*

Introduction

One of the important proteins associated with water stress is Late Embryogenesis Abundant (LEA) proteins. LEA proteins accumulate in vegetative plant tissues at the commencement of seed desiccation and response to water deficiency. It has been divided into distinct families, each with its own set of conserved motifs (Carillo, 2011). Many members of the LEA family can be found in higher plants. They can be split into eight subgroups based on amino acid sequence similarities and differences in conserved domains (LEA1, LEA2, LEA3, LEA4, LEA5, LEA6, dehydrin [DHN], and seed maturation protein [SMP]). These proteins are important in tolerance to the water stress which leads to desiccation and cold shock. It has been proposed that LEA proteins may act as molecular chaperones which prevent the formation of destructive protein accumulation at the time of water stress (Goyal, 2005). Since this family of proteinsfinds a great place in water stress, classifying and predicting the family of proteins is important. With the influx of massive volumes of genetic data, extremely precise protein identification and categorization have become critical. Some protein studies have used machine learning approaches in recent years, with encouraging findings. For this classification problem, Support vector Machine (SVM) classification will be greatly helpful. SVM, Logistic Regression, Random forest, and Convolutional Neural Network algorithm were used to classify the LEA proteins concerning their

([1] PG Scholar, [2]Associate Professor (Maths), [3]Associate Professor (Computer Science), [4]Assistant Professor (Bioinformatics)

hydrophobic characteristics (Hu, 2020).

Materials and Methods

Data collection

The protein sequences were collected from LEAPDb, a database for the late embryogenesis abundant proteins (Hunault, 2010), and additional sequences were collected from National Centre for Biotechnology Information (NCBI) and Uni Prot. The negative proteins collected were does not belong to any of the classes of positive proteins. The number of proteins belonging to each class were listed in Table 1.

Table 1. Number of proteins belonging to a positive and negative dataset

Sub-family	Positive	Negative
LEA-1	85	107
LEA-2	399	530
LEA-3	76	130
LEA-4	282	296
LEA-5	70	98
LEA-6	146	171
Dehydrin	492	513
SMP	64	88

Feature Extraction from Protein Sequence

The classification of proteins was done based on the physico-chemical properties of the proteins. The Prot Param tool is used for this purpose. The various physico-chemical properties like theoretical pI, molecular weight, atomic composition, extinction coefficient, instability index, aliphatic index, and grand average of hydropathicity (GRAVY). Some of the properties like bulkiness, hydrophobicity, fold Index, transmembrane tendency, and Average area buried, net charge, and mean hydrophilicity are extracted with the help of the Prot Scale tool.

Support vector machine

Support Vector Machine is a prominent Supervised Learning technique that can be used to solve both classification and regression problems. The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space into classes so that additional data points can be readily placed in the correct category in the future. The linear kernels are used when the data is linearly separable and non-linear kernels are used when the data is non-linearly separable. The non-linear kernels include polynomial kernel, Gaussian RBF kernel, and sigmoid kernel. In this study polynomial and RBF kernels were used with default parameters. The SVM module was performed using the weka 3.8.6 toolkit.

Performance evaluation

In this study, the performance of our model was measured using specificity, sensitivity, accuracy, Root Mean Square Error (RMSE), and Mathew correlation coefficient (MCC). Also, the model was evaluated using 5-fold cross-validation techniques. A confusion matrix is a consolidated table that is used to evaluate a classification model performance. The confusion matrix consists of four components namely true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).The accuracy, precision, and recallof an algorithm are performance metrics that are derived based on TP, TN, FP, and FN (Singh *et al.,* 2021). Precision is measured as the ratio of correctly classified samples as the positive samples (TP) to the total number of positive samples (TP+FP). Recall metric also called sensitivity is defined as the ratio of correctly classified positive samples (TP) divided by a total number of positive samples. MCC is a more accurate statistical rate that yields a high score only if the forecast is correct in all four areas of the confusion matrix (Chiggo, 2020). The RMSE is a performance measure and tells how far awaythe model is from being correct. In k-fold cross the dataset should be first divided into k equally (or nearly equally) sized segments or folds. Following that, k iterations of training and validation are carried out, with each iteration holding out a different fold of the data for validation and the remaining k-1 folds being used for learning. (Refaeilzadeh *et al.,* 2009).

Results and Discussion

In this study, the total dataset was divided into training data and test data in a ratio of 70:30. Eight binary SVM-based modules were developed for the discrimination of LEA proteins from other proteins by physico-chemical properties using polynomial and RBF kernel with default parameters. The performance of different kernels were compared using different metrics and displayed in Table 2 and it is evident that the accuracy of the polynomial kernel is better than the RBF kernel for all classes except for

SMP class. The class LEA-1 has the highest accuracy of 97.1 per cent. The accuracy of the polynomial kernel ranges from 89.1 to 97.1 per cent and for the RBF kernel, it ranges from 64.5 to 95.7 per cent. The sensitivity of polynomial kernel ranges from 86.7 to 100 per cent, whereas for RBF kernel it ranges from 58.4 to 100 per cent. The model with 100 per cent sensitivity indicates that it has correctly predicted all the true positives. Furthermore, in the case of specificity, the polynomial kernel ranges from 83.9 to 100 per cent, and for RBF kernel ranges from 84.3 to 100 per cent.

The model performance evaluated using 5-fold cross validation techniques were also tabulated in Table 2.

The MCC value ranges from 0.6 to 0.9 which is a good measure. For LEA-1, 2, 4, 5, 6, and dehydrin classes, the precision, recall, and MCC of the polynomial kernel are better than the RBF kernel. For the class of LEA-3, both the kernels perform almost similarly. For the SMP class, the RBF performs better than the polynomial kernel. The highest recall measure were recorded for polynomial kernel for LEA-5 class indicates that the model can recall 96 per cent of proteins correctly.

In Fig. 1 the RMSE values for the 5-fold cross-validation for each of the kernels have been displayed. The results show that the error of the polynomial kernel is less than the RBF kernel in all the class of LEA proteins except the class LEA-3 where the error per cent of the RBF kernel is less than the polynomial. The overall performance of the polynomial kernel is better than the RBF kernel.

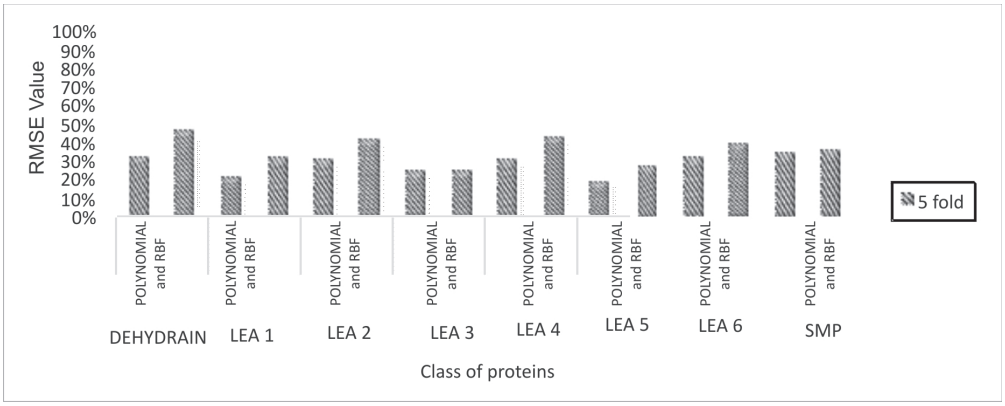

Fig. 1. RMSE values for different cross-validation techniques

Table 2. Performance of SVM modules in the classification of LEA proteins

Protein class	Kernel function	Sensitivity %	Specificity %	Accuracy %	5-fold cross validation		
					Precision	Recall	MCC
DEHYDRIN	Polynomial	86.7	93.7	90.0	0.905	0.898	0.803
	RBF	58.4	84.3	64.5	0.818	0.791	0.609
LEA 1	Polynomial	96.4	97.5	97.1	0.959	0.958	0.917
	RBF	96.4	95.1	95.7	0.919	0.906	0.825
LEA 2	Polynomial	91.3	91.4	91.4	0.911	0.911	0.818
	RBF	76.5	98.2	85.3	0.877	0.832	0.710
LEA 3	Polynomial	92.0	100.0	96.8	0.948	0.947	0.888
	RBF	100.0	92.9	95.2	0.947	0.942	0.878
LEA 4	Polynomial	95.1	89.1	91.9	0.915	0.913	0.828
	RBF	72.7	88.9	78.6	0.831	0.822	0.653
LEA 5	Polynomial	95.0	93.9	94.3	0.966	0.966	0.929
	RBF	87.0	96.7	92.5	0.933	0.927	0.855
LEA 6	Polynomial	88.9	89.8	89.5	0.902	0.902	0.803
	RBF	74.4	88.5	82.1	0.861	0.858	0.718
SMP	Polynomial	100.0	83.9	89.1	0.885	0.882	0.763
	RBF	83.3	100.0	91.3	0.904	0.875	0.778

Conclusion

In this work, we have developed an SVM model for the LEA proteins classification model including polynomial and RBF kernel. The confirmed LEA proteins were used as positive samples of eight different classes in the model training. Different indicators are used for model comparisons.The estimates of errors were derived using 5-fold cross-validation procedures and the models' comparative performances were tested using various criteria such as sensitivity, accuracy, MCC, and specificity. RMSE values for the different cross-validation were also compared. The overall performance of the polynomial kernel was found to be better than the performance of the RBF kernel. Only in the case of class LEA 3 the performance of the RBF kernel is found to be slightly better than the polynomial kernel. These developed models could be used for further studies on LEA proteins and the development of a web-based server which would greatly reduce time and cost.

References

Chicco, D. and Jurman, G. 2020. The advantages of the Matthews correlation coefficient (MCC) over the F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 21 (1): 1-13.

Goyal, K., Walton, L. J. and Tunnacliffe, A. 2005. LEA proteins prevent protein aggregation due to water stress. *The Biochemical Journal*. 388 (1): 151–157.

Hu, S., Zhao, M., Shi, Z. and Zhang, M. 2020. Prediction of LEA Plant Proteins Based on Machine Learning. In: *Journal of Physics: Conference Series*. 1631 (1).

Hunault, G. and Jaspard, E. 2010. LEAPdb: a database for the late embryogenesis abundant proteins. *BMC Genomics.* 11(1) : 1-9.

Olvera-Carrillo, Y., Luis Reyes, J. and Covarrubias, A. 2011. Late embryogenesis abundant proteins: versatile players in the plant adaptation to water limiting environments. *Plant Signaling & Behavior*. 6 (4) : 586-589.

Refaeilzadeh, P., Tang, L. and Liu, H. 2009. Cross-validation. *Encyclopedia of Database Systems*. 5 : 532-538.

Singh, P., Singh, N., Singh, K. K. and Singh, A. 2021. Diagnosing of disease using machine learning. In : *Machine Learning and the Internet of Medical Things in Healthcare*. 89-111.