International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 99-105
e-ISSN:2278-621X

A MACHINE LEARNING TOOL FOR SERK PROTEIN USING SVM

Anusha T.A[1], Ashraf K.M[2] and Hemalatha N[3]

Abstract- The process where a plant or embryo is derived from single somatic cells is known as Somatic embryogenesis. Somatic Embryogenesis Receptor Kinase (SERK) is a protein having direct role in Somatic Embryogenesis. A support vector machine (SVMs, also support vector networks) is an algorithm in machine learning that evaluate data used for classification and regression analysis. In this paper we have used SVM, supervised learning model to predict SERK proteins. The input data for developing the same was taken from NCBI which were SERK proteins in palm plants. The machine is trained to identify the SERK protein sequence. Best predictive model was obtained for N terminal 15 and 20 composition where an accuracy of 80% was obtained.A web server for this predictive model was implemented.

Keywords – Machine learning, SERK, SVM, NCBI, N terminal,C terminal.

I. INTRODUCTION

Somatic embryogenesis is a process where a plant or embryo is derived from single somatic cells. Somatic embryos are developed from plant cells that are not commonly involved in the development of embryos that is normal plant tissue. For asexual propagation or somatic cloning in plants somatic embryogenesis is a unique pathway [1, 2].

In Somatic Embryogenesis SERK Protein has its direct role. In Cell metabolism which leads to growth and defense responses, exogenous signals play an important role. Some of these stimuli activate anatomical and physiological modifications that are normally modulated by gene expression. *SERK*belongs to a small family of genes that code for a transmembrane protein involved in signal transduction and that have been strongly combined with somatic embryogenesis and apomixis in a number of plant species. The latest studies authenticate its role in somatic embryogenesis and recommend an extensive range of functions in plant response to biotic and abiotic stimuli.

Biochemical and morphological changes occur throughout the development of induced tissues in the course of somatic embryogenesis, which is closely related to alterations in gene expression. Certain genes are differentially expressed during somatic embryogenesis induction, at the same time others are expressed during differentiation from embryo maturation up to full plant development.*SERK* gene is declared to have a considerable role in the induction of somatic embryogenesis.

SERK is a leucine-rich repeat (LRR) transmembrane protein kinase that reinforce the ability of the apical meristem in *Arabidopsis* to form somatic embryos [3] LRR kinases form homodimers or heterodimers with other receptor-like kinases (RLKs), in response to binding by a ligand to transmit their signal. This ligand-induced dimerization causes phosphorylation of the intracellular kinase domains of the RLKs, which trigger the next stages of the signal transduction pathway. There is potential for

[1] *Aloysius Institute of Management and Information technology, Mangalore,Karnataka , India.*
[2] *Aloysius Institute of Management and Information technology, Mangalore,Karnataka , India.*
[3] *Aloysius Institute of Management and Information technology, Mangalore,Karnataka , India.*

different levels of complexity in the signaling through variation in the binding partners of different RLKs [4].

In this study we have collected the list of SERK nucleotide sequence from NCBI and collected the protein sequence from the ORF finder. These sequences were used to create predictive model using SVM. The predictive system was trained using three kernels of svm - linear, polynomial and RBF. These kernels were trained to identify the sequence based on the N terminal, C terminal and full length of the given protein sequence. The predictive result of the same has been discussed in the result section.

II.MATERIALS ANDMETHODS

A. Dataset

Nucleotide sequences representing the SERK proteins for different plants were collected from NCBI. List of plants that contain SERK Protein are as follows:

Momordicacharantia,Dendrobiumcatenatum,Larix decidua,Curcuma alismatifolia,Heveabrasiliensis,Cyrtochilumloxense,Theobroma cacao,Dimocarpuslongan,Lactuca sativa,Coffeacanephora,Citrus unshiu,Anthuriumandraeanum,Citrus sinensis,Cattleya maxima,Poapratensis,Cyclamen persicum ,Oryza sativa,Zea mays, Helianthus annuus ,Daucuscarota,Arabidopsis thaliana ,Camellia nitidissima,Medicagotruncatula, Gossypiumhirssutum ,Triticumaestivum ,Ananascomosus ,Anthuriumandraeanum.

Reference protein sequences were taken from the ORF finder. 51 Amino acid sequences were obtained. ORF Finder is a tool in Bioinformatics which searches for open reading frames (ORFs) in the DNA sequence. The program returns the range of each ORF, along with its protein sequence.An Open reading frame (ORF) in molecular genetics is the region of a reading frame that has the capacity to code for a protein or peptide. An ORF is a unbroken section of codons that do not contain a stop codon (usually UAA, UAG or UGA) [5].The transcription termination site is placed after the ORF, ahead the translation stop codon, During translationan incomplete protein would be formed, if transcription were to drop before the stop codon [6].

In this paper train set was created using 41 Positive and 41 negative amino acid sequences.Positive sequences were taken from the above datasets;negative sequences were taken from the NCBI which is non SERK Protein.10 Positive and 10 negative amino acid sequence were used to create the test set. Positive sequences were taken from the datasets;negative sequences were taken from the NCBI which is non SERK Proteins.

B. Features

In this work, three features were used. For finding the features in the sequence perl program was used. In the Amino acid composition count of existence of each amino acid in the sequence was observed and was calculated. In the N terminal composition, 3 different lengths *viz.*15, 20, 25 and for C terminal 15, 20, 25 was considered. To calculate the count of amino acid in each protein sequence, following formula is used:

$$\text{Fraction of ith amino acid} = \frac{\text{Total number amino acid i}}{\text{Total number of amino acid in protein}} \quad (1)$$

Perl programming language was used to generate the test and train sets.Perl is a general-purpose programming language basically developed for text manipulation.

C. SVM

In this study, predictions with classification method was estimated using a strong machine learning technique,Support vector machine(SVM). SVM, a machine learning method, has been utilized for many kinds of pattern recognition problems.The principle of the SVM is to convert the samples into a higher dimension space called Hilbert space. A separating hyper plane is sought in this space called the optimal separating hyperplane in such a way as to maximize its distance from the closest training samples. SVM is a supervised machine learning technology rooted theoretically on statistical learning theory [7].In this work to implement SVM,SVM light package was used,which allows to choose number of parameters and kernels(eg:linear,Polynomial, and radial basis function).Linear kernel is often recommended for text classification.The selection of kernel plays asignificant role in SVM and is related to selecting architecture in artificial neural network.In this paper,learning was done using three kernels:linear,Polynomial, and radial basis function.

D. Classification measure

In the criteria of classification of a prediction, the sequenceswere divided into true positive(TP),true negative(TN),false positive(FP),or false negative(FN). True positive prediction means positive sample is predicted as positive, it is classified under true positive. If negative sample is predicted as negative then it is classified under true negative.If a positive sample is predicted as negative class and vice versa then it is classified as false negative and false positive prediction,respectively.
To assess the performance of gene prediction tool,the standard prediction measures by Burset and Guigo were applied [8].Description of these parameters is given below:

i. Sensitivity:It gives the amount of SERK gene correctly predicted.
ii. Specificity:It gives the amount of non SERK gene correctly predicted.
iii. Accuracy:It gives the total number of predictions that were correct.
iv. Precision:It is the proportion of the predicted positive cases that were correct.
v. F measure:It commonly used for “average” of precision and sensitivity

These parameters can be calculated using following equations

$$Sensitivity = (TP)/(TP + FN) \quad (2)$$

$$Specificity = (TN)/(FN + TN) \quad (3)$$

$$Accuarcy = (TP + TN)/(TP + TN + FP + FN) \quad (4)$$

$$Precision = TP/(TP + FP) \quad (5)$$

$$F\ measure = (Precision)(Sensitivity)/(Precision + Sensitivity) \quad (6)$$

Where TP and TN are correctly predicted SERK and non SERK gene,respectively.FP and FN are wrongly predicted SERK and non SERK genes respectively. matthews correlation coefficient (MCC) was used to measure the quality of binary (two-class) classifications in machine learning[9]. It takes true and false positives and negatives and is generally remarked as a balanced measure which can be used even if the classes are of very different sizes. The MCC returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 represents total disagreement between prediction and observation.Positive MCC value stands for better prediction performance.MCC is calculated using the equation 7[10].

$$MCC = TP \times TN - FP \times FN/\sqrt{(TP + FP)(TP + FN)(TN + FN)} \quad (7)$$

For estimating the performance of the prediction tool, independent data test and cross validation test were carried out. In independent data test validation, training dataset and testing set is considered to be

independent of one another and hence the name. Cross validation is a model validation method for measuring how the results of a statistical analysis will generalize to an independent data set.

III. RESULT

The prediction accuracy of the SVM based classifier was assessed by two distinct approaches: cross validation test and independent dataset tests. Testing of SVM on independent dataset for SERK Protein in various plants achieved a maximum accuracy of 80% with an MCC value of 0.6 using linear kernel for N terminal composition with window size of 20 and 15.Here sensitivity was 0.7 and specificity 0.9 for N terminal 20 and 0.8 the sensitivity and specificity for N terminal 15 (Table 1) .Hence it can be concluded that for independent data set test, N terminal 15 or 20 is optimal with linear kernel as classifier.

Testing of SVM on cross validation for SERK Protein achieved an accuracy of 80% using linear kernel for N terminal composition with window size of 15. MCC value for this was 0.6 with sensitivity and specificity0.8. This result was better compared to other compositions (Table 2). Hence it can be concluded that for cross validation set test, N terminal 15 was optimal with linear kernel as classifier.

Table-1SVM results of Independent Data set test

Compostion	Kernel	Sensitivity	specificity	Accuarcy	F measure	Precision	MCC
Nterm15	Linear	0.8	0.8	0.8	0.4	0.8	0.6
	Poly	0.5	1	0.75	0.333	1	0.577
	Rbf	0.1	1	0.55	0.05	0.1	0.333
Nterm20	Linear	0.7	0.9	0.8	0.388	0.875	0.612
	Poly	0.5	0.8	0.65	0.294	0.714	0.314
	Rbf	0.1	1	0.55	0.09	1	0.229
Nterm25	Linear	0.9	0.5	0.7	0.428	0.818	0.436
	Poly	0.7	0.9	0.8	0.388	0.875	0.612
	rbf	0.1	1	0.55	0.09	1	0.229
Cterm15	Linear	0.6	0.7	0.65	0.312	0.666	0.229
	Poly	0.7	0.7	0.7	0.35	0.7	0.4
	Rbf	0.5	1	0.75	0.333	1	0.577
Cterm20	Linear	0.7	0.4	0.55	0.318	0.583	0.149
	Poly	0.8	0.7	0.75	0.38	0.727	0.479
	Rbf	0.3	1	0.65	0.23	1	0.42
Cterm25	Linear	0.9	0.5	0.7	0.374	0.642	0.377
	Poly	0.9	0.7	0.8	0.409	0.75	0.612
	Rbf	0.1	1	0.55	0.09	1	0.229
Fulllength	Linear	0.4	0.9	0.65	0.266	0.8	0.346
	Poly	0.4	0.9	0.65	0.266	0.8	0.346
	Rbf	0.1	1	0.55	0.09	1	0.229

Graphical representation of SERK predictor using N terminal composition with length 15, 20 and 25 using linear kernel of SVM are represented in Figure 1.

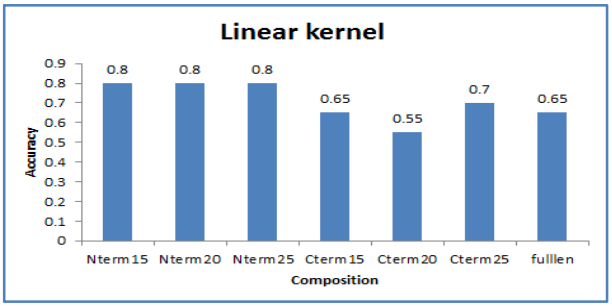

Figure 1.Performance Chart of accuracy for different composition methods for linear kernel for the Independent data set test

Table 2 SVM results of Cross Validation Data set test

Compostion	Kernel	Sensitivity	specificity	Accuracy	F measure	Precision	MCC
Nterm15	Linear	0.8	0.8	0.8	1	0.444	0.6
	Poly	0.6	1	0.8	1	0.375	0.654
	Rbf	0.1	1	0.55	1	0.09	0.229
Nterm20	Linear	0.7	0.9	0.8	1	0.411	0.612
	Poly	0.5	0.8	0.65	0.714	0.294	0.688
	Rbf	0.1	1	0.55	1	0.09	0.241
Nterm25	Linear	0.6	1	0.8	1	0.375	0.561
	Poly	0.6	1	0.8	1	0.375	0.654
	rbf	0.1	1	0.55	1	0.09	0.316
Cterm15	Linear	0.6	0.7	0.65	0.666	0.315	0.301
	Poly	0.7	0.7	0.7	0.7	0.35	0.326
	Rbf	0.5	1	0.75	1	0.333	0.577
Cterm20	Linear	0.7	0.4	0.55	0.538	0.304	0.281
	Poly	0.8	0.7	0.75	0.727	0.38	0.502
	Rbf	0.3	1	0.65	1	0.75	0.42
Cterm25	Linear	0.9	0.5	0.7	0.642	0.411	0.436
	Poly	0.9	0.7	0.8	0.75	0.409	0.548
	Rbf	0.1	1	0.55	1	0.09	0.229
fulllength	Linear	0.4	0.9	0.65	0.8	0.266	0.346
	Poly	0.4	0.9	0.65	0.8	0.266	0.346
	Rbf	0.1	1	0.55	1	0.266	0.229

SERK PROTIEN PREDICTION is a website which we developed to implement the prediction of serk protein. Here home page displays all the menus and description of serk [Figure 4]. In the tool page we can input the sequence and user will get the output[Figure 5].

Figure 4: Home page

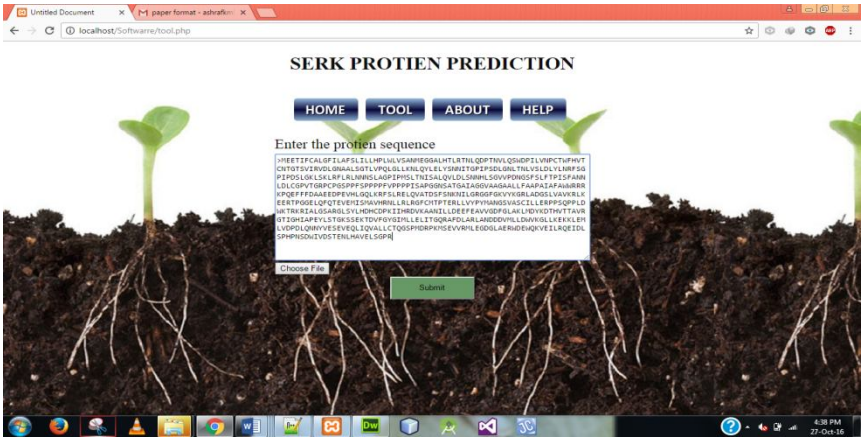

Figure 5: Tool page

Figure 6: Output page

IV CONCLUSIONS

In this paper we have worked out SERK predictor model using SVM classifier. The three kernels *viz*., linear, polynomial and RBF kernels were applier modeled for three compositions. Performance evaluation of the predictor model was done through independent test and cross validation test. It was

observed that for both test, N terminal composition with window length 15 obtained a better result with accuracy 80%. A web server was also developed for the predictive model. This kind of predictive model would be really beneficial for biologists for future annotations of novel Palm proteins.

REFERENCES

[1] http://www.accessexcellence.org/LC/ST/st2bgplant.htmlPlant Tissue Culture.

[2]. George, Edwin F., Hall, Michael A., De Klerk, Geert," Somatic Embryogenesis", Plant Propagation by Tissue Culture,vol 1 ,pp.335-354, 2008.

[3]. Hecht V, Vielle-Calzada JP, Hartog MV, Schmidt ED,Boutilier K and Grossniklaus U, "The Arabidopsis SOMATIC EMBRYOGENESIS RECEPTOR KINASE 1 gene is expressed in developing ovules and embryos and enhances embryogenic competence in culture",Plant Physiol, vol127(3), pp803–816,2001.

[4]. Nolan KE, Kurdyukov S ,Rose RJ , "Expression of the SOMATIC EMBRYOGENESIS RECEPTOR-LIKE KINASE1 (SERK1) gene is associated with developmental change in the life cycle of the model legume Medicagotruncatula."Journal of Experimental Botany,vol60(6), pp1759–1771 ,2009.

[5]. http://www.britannica.com/EBchecked/topic/1116194/machine-learning

[6]. Phil Simon,Too Big to IgnoreThe Business Case for Big Data. Wiley.pp. 89.ISBN 978-1-118-63817-0,2013

[7]. Ron Kohavi; Foster Provost ."Glossary of terms".Macine Learningvol30,pp 271–274,1998.

[8]. Machine learning and pattern recognition "can be viewed as two facets of the same field."

[9]. Wernick, Yang, Brankov, Yourganov and Strother. "Machine Learning in Medical Imaging"IEEE Signal Processing Magazine, vol. 27, no. 4, pp. 25-38, 2010

[10]. Russell, Stuart, Norvig and Peter , "Artificial Intelligence: A Modern Approach (2nd ed.)."Prentice Hall.ISBN 978-0137903955,2003 Open reading frame. U.S. National Library of Medicine (2015).

[11]. Slonczewski, Joan, John Watkins Foster An Evolving Science. Microbiology New York: W.W. Norton & Co. ISBN 978-0-393-97857-5. OCLC185042615,2009

[12]. Hua S and Sun Z ,"A novel method of protein secondary structure prediction with high segment overlap measure support vector machine approach", J MolBiolVol 308,pp:397–407,2001

[13]. ScholkopfB,Burges C &Samola A T Joachims,"Advances in kernel Methods –support Vector Learning"MIT Press,Cambridge,MA,1999

[14]. Matthews, B. W. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". Biochimica et BiophysicaActa (BBA) - Protein StructureVol405 (2),pp: 442–451, 1975.

[15]. Powers, David M W ,"Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness& Correlation" ,Journal of Machine Learning Technologies. Vol2 (1),pp 37–63,2011.